DOCUMENT RESUME

ED 292 862 TM 011 268

AUTHOR	Kavanagh, Michael J.; And Others
TITLE	Job Performance Measurement in the Military: A Classification Scheme, Literature Review, and Directions for Research. Final Technical Report for Period May 1982-February 1983.
INSTITUTION	McFann-Gray and Associates, Monterey, CA.
SPONS AGENCY	Air Force Human Resources Lab., Brooks AFB, Tex. Manpower and Personnel Div.
REPORT NO	AFHRL-TR-87-15
PUB DATE	Sep 87
CONTRACT	F41689-81-C-0022
NOTE	68p.
PUB TYPE	Information Analyses (07J) -- Reports - Descriptive (141) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Classification; Evaluation Criteria; *Job Performance; Measurement Techniques; *Military Personnel; Occupational Tests; Officer Personnel; Performance Tests; *Personnel Evaluation; Personnel Selection; Rating Scales
IDENTIFIERS	*Air Force Human Resources Laboratory

ABSTRACT

A classification scheme for job performance measurement was developed by the Air Force Human Resources Laboratory (AFHRL) for development of individual job performance measures to validate Air Force selection and classification procedures. The job performance literature indicates that previous research relied heavily on broad-based generic indices, performance ratings, or operational measures with their inherent problems of inflation and halo effects. These broad measures were unable to take into account task-level-specific influences such as training differences or opportunities to perform. However, it appears that current interest, resources, and state-of-the-art technology developments have significantly increased the probability of developing successful measures of job performance. The AFHRL study resulted in guidelines for developing and obtaining performance measures for a wide range of enlisted, officer, and civilian jobs. Research implications are associated with individual characteristics, rater-ratee relationships, measurement methods and scale development, scale characteristics, performance standards, social contexts, non-work variables, performance constraints, organizational norms, public relations and administrative procedures, rater training, intervening variables, measurement and research paradigms, and error variance. A 12-page list of references is provided. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

AFHRL-TR-87-15

AIR FORCE HUMAN RESOURCES LABORATORY

JOB PERFORMANCE MEASUREMENT IN THE MILITARY: A CLASSIFICATION SCHEME, LITERATURE REVIEW, AND DIRECTIONS FOR RESEARCH

Michael J. Kavanagh
School of Business
State University of New York at Albany
Albany, New York 12222

Walter C. Borman
Personnel Decisions Research Institute
43 Main Street, S.E.
Suite 405
Minneapolis, Minnesota 55414

Jerry W. Hedge

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

R. Bruce Gould

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

September 1987
Final Technical Report for Period May 1982 - February 1983

Approved for public release; distribution is unlimited.

ED292862

TM 011 268

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

BEST COPY AVAILABLE

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

HENDRICK W. RUCK, Technical Advisor
Training Systems Division

HAROLD G. JENSEN, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5 MONITORING ORGANIZATION REPORT NUMBER(S) AFHRL-TR-87-15
6a. NAME OF PERFORMING ORGANIZATION McFann - Gray and Associates 6b OFFICE SYMBOL (If applicable)	7a NAME OF MONITORING ORGANIZATION Training Systems Division
6c. ADDRESS (City, State, and ZIP Code) 2100 Garden Road, Suite J Monterey, California 93940	7b ADDRESS (City, State, and ZIP Code) Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235-5601
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Human Resources Laboratory 8b OFFICE SYMBOL (If applicable) HQ AFHRL	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F41689-81-C-0022
8c. ADDRESS (City, State, and ZIP Code) Brooks Air Force Base, Texas 78235-5601	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 62703F; PROJECT NO: 7719; TASK NO: 18; WORK UNIT ACCESSION NO: 21
11. TITLE (Include Security Classification) Job Performance Measurement in the Military: A Classification Scheme, Literature Review, and Directions for Research	
12. PERSONAL AUTHOR(S) Kavanagh, M.J.; Borman, W.C.; Hedge, J.W.; Gould, R.B.	
13a. TYPE OF REPORT Interim 13b. TIME COVERED FROM May 82 TO Feb 83	14. DATE OF REPORT (Year, Month, Day) September 1987 15 PAGE COUNT 68
16. SUPPLEMENTARY NOTATION	

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	criterion development job performance assessment performance measurement
05	09		
05	10		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The major purpose of this report is to describe a classification scheme of job performance measurement, with emphasis on its applicability in a military context. This report is organized into five sections: (a) an introduction, which provides report organization and background; (b) a description of a conceptual performance measurement classification scheme used to organize and categorize research findings; (c) an examination of the empirical and theoretical literature relevant to the variables and relationships identified in the schema; (d) specific recommendations for both applications and research directions for a long-term R&D effort; (e) priorities related to research directions, both from an importance and urgency standpoint.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Nancy J. Allin, Chief, STINFO Office	22b TELEPHONE (Include Area Code) (512) 536-3877 22c OFFICE SYMBOL AFHRL/TSR

DD Form 1473, JUN 86 *Previous editions are obsolete.*

SECURITY CLASSIFICATION OF THIS PAGE

Unclassified

SUMMARY

The job performance measurement literature indicates that previous research relied heavily on broad-based generic indices, performance ratings, or operational measures with their inherent problems of inflation and halo effects. These broad measures were unable to take into account task-level-specific influences such as training differences or opportunities to perform; hence, such efforts have been largely unsuccessful. However, it appears that current interest, resources, and state-of-the-art technology developments have now significantly increased the probability of developing successful measures of job performance. This report describes the Air Force Human Resources Laboratory's (AFHRL) research program for development of individual job performance measures. The report describes the construction of a job performance measurement classification scheme into which the relevant empirical and theoretical literature are organized. Based on this framework, specific recommendations for both applications and research directions are given.

PREFACE

This report describes the initiation of a long-term program of research and development (R&D) focusing on job performance criterion development. The work was performed by McFann-Gray and Associates, Inc., under contract F41689-81-C-0022 with the Air Force Human Resources Laboratory (AFHRL), Manpower and Personnel Division. The work was accomplished under Work Unit 77191821. Dr. R. Bruce Gould was the AFHRL Contract Monitor.

Several influences have highlighted the Air Force's need for performance measurement and brought ongoing and planned programs to their current state. Planning for the research program began several years ago on the recommendation of two Research Advisory Panels (composed of knowledgeable scientists from academia and industry, as well as peers from the Army and Navy). They reviewed the entire AFHRL manpower, personnel, and training research program and recommended consolidation of separate measurement efforts into one unified research program. At the same time, the Uniform Guidelines for Employee Selection (1978) and a review of case law mandated that Air Force civilian selection systems be validated against job performance measures. Finally, Congress mandated that military selection tests be validated against hands-on job performance measures. These operational, legal, and Congressional mandates have thus provided the impetus to planning and obtaining support for a lengthy, high resource research effort.

The short-term objective of this effort is the development of on-the-job performance measures to validate Air Force selection and classification procedures. Guidelines for developing and obtaining the performance measures will be established for a wide range of enlisted, officer, and civilian jobs. Once obtained, the measures will be placed in a data base for validation use. The long-term goal is to establish an operational performance measurement program for evaluation of selection and training procedures, as well as personnel policies and practices. The goal here is to operationalize the procedures so that performance measurement and evaluation can be carried on by technicians, as is currently done by the USAF Occupational Measurement Center with the Occupational Survey (Job Analysis) Program. In this way, R&D resources will be freed for other projects.

TABLE OF CONTENTS

		Page
I.	INTRODUCTION	1
	1.1 Organization of Report	1
	1.2 Terminology	2
	1.3 Background	3
	1.4 Classification Scheme Boundaries	4
II.	A CONCEPTUALLY BASED DESCRIPTIVE CLASSIFICATION SCHEME OF PERFORMANCE MEASUREMENT QUALITY	5
	2.1 A Simplified Job Performance Schema	5
	2.2 Development of a General Job Performance Measurement Classification Scheme	7
	2.3 Validation Research Classification Scheme	12
	2.4 Additional Considerations for the AFHRL Program	15
III.	LITERATURE REVIEW	15
	3.1 Introduction	15
	3.2 Individual Characteristics and Measurement Quality	16
	3.3 Rater-Ratee Relationship	17
	3.4 Measurement Method	19
	3.5 Performance Standards, Scale Characteristics, and Scale Development	21
	3.6 Environmental Context, Non-Work Variables, Performance Constraints, and Organization/Unit Norms	24
	3.7 Public Relations and Administrative Procedures	24
	3.8 Rater Training	25
	3.9 Intervening Process Variables	26
IV.	RESEARCH IMPLICATIONS	26
	4.1 Individual Characteristics	27
	4.2 Rater-Ratee Relationships	28
	4.3 Measurement Method	28
	4.4 Measurement Scale Development	30
	4.5 Scale Characteristics	30
	4.6 Performance Standards	31
	4.7 Social Context	31
	4.8 Non-Work Variables	32
	4.9 Performance Constraints	32
	4.10 Organization/Unit Norms	33
	4.11 Public Relations/Administrative Procedures	33
	4.12 Rater Training	33
	4.13 Intervening Variables	34
	4.14 Measurement and Research Paradigm Issues	34
V.	RESEARCH PRIORITIES	35
	5.1 Measurement Methodology	36
	5.2 Scale Development and Characteristics	36

Table of Contents (Concluded)

Page

5.3 Research Paradigm Issues 37
5.4 Identification of Possible Sources of Error Variance 38
5.5 The Control of Error Variance 38
5.6 Final Comments 39

REFERENCES 40

APPENDIX A: RESEARCH ISSUES 53

LIST OF FIGURES

Figure Page

1 A Simplified Job Performance Scheme 6

2 A Job Performance Measurement Classification Scheme 7

3 A Job Performance Measurement Classification Scheme for Validation Research 12

LIST OF TABLES

Table Page

1 Quality of Performance Measurement Criteria 7

2 Variables That Can Impact on Measurement Quality 10

JOB PERFORMANCE MEASUREMENT IN THE MILITARY: A CLASSIFICATION SCHEME, LITERATURE REVIEW, AND DIRECTIONS FOR RESEARCH

I. INTRODUCTION

The major purpose of this report is to describe a job performance measurement classification scheme, with emphasis on its applicability in a military context. The field of job performance measurement has probably generated more literature in the behavioral sciences than has any other topic, yet there does not yet exist a complete conceptual framework for this phenomenon. The works of DeCotiis and Petit (1978) and Wherry and Bartlett (1982) represent the most significant efforts at providing partial conceptual frameworks, and these will be reviewed in more detail later in this report. The importance of these two conceptualizations to this effort is that they share the same perspective that accuracy of the performance evaluation is the most critical indicator of the quality of the measurement.

The lack of a complete conceptual framework for the measurement of job performance is central to the problem of properly specifying and measuring dependent variables in both causal and covariate designs. This is particularly problematic for the applied researcher who is concerned with understanding and predicting the behavior of people in organizations. Within the military context, this problem becomes more acute for scientists involved in research and recommendations for action in any of the traditional personnel decision functions. Thus, a major outcome of this report will be a conceptually based descriptive classification scheme of performance measurement variables that may be used (a) to summarize and organize research progress in terms of previous empirical work and (b) to identify future research and development (R&D) needs. These two outcomes should prove helpful to the long-term R&D program being initiated by the Air Force Human Resources Laboratory (AFHRL) to develop a methodology for measuring job performance in the military.

1.1 Organization of Report

The four chapters that follow will provide: (a) a description of a conceptual performance measurement classification scheme; (b) an examination of the empirical and theoretical literature relevant to the variables and relationships identified in the schema; (c) specific recommendations for both applications and research directions; and (d) priorities related to research directions.

The second chapter is an integration and, by necessity, a deductive extension of previous attempts to provide conceptual descriptions of parts of the performance measurement situation (see for example, Cummings & Schwab, 1973; DeCotiis & Petit, 1978; Kavanagh, 1982a; Landy & Farr, 1980; MacKinney, 1967; Ronan & Prien, 1971; Wherry & Bartlett, 1982). The integration of previous conceptualizations is necessary because none completely describes all aspects of the performance measurement situation as envisioned in this report. The second chapter provides a conceptually based descriptive classification scheme that serves as a mechanism for organizing the literature review and prescribing needed research.

The third chapter examines the literature on performance measurement. Computer searches of both the public (e.g., Psych SCAN) and Department of Defense (DOD) literature were conducted to identify as much of the relevant literature as possible. In addition, behavioral scientists identified with the performance measurement literature were contacted in an attempt to uncover current, unpublished studies related to this topic. This chapter represents a first attempt to verify the relationships or linkages hypothesized in the classification scheme and provides a

summary showing the empirical support (or non-support) for these relationships. This literature review serves as a basis for revision of the schema and provides direction for AFHRL's program of research.

The fourth and fifth chapters are most important, in that they can be used as guides for the long-term R&D program within AFHRL to develop a measurement methodology for job performance.

The fourth chapter, organized by the linkages in the model, contains recommendations both for specific features to include in the design of the measurement methodology and for specific areas where research is needed. The recommendations will help to conserve AFHRL resources by also specifying where research is not necessary (i.e., where prescriptive advice exists in the literature).

The final chapter provides recommendations for research that are prioritized in terms of their importance to the overall program of R&D at AFHRL, presented in chronological order to serve as a planning tool, and integrated within the conceptually based classification scheme in this report.

1.2 Terminology

Before proceeding further, it is important to define and differentiate among the various terms used in the field of performance measurement. Criterion, one of the most commonly used terms in the field, refers to a measure of performance. In the context of this report, a criterion is a measure of an individual's performance on a job. Performance measure is essentially the same as criterion for the purposes of this report, and these two terms will be used interchangeably; however, it is possible to have more than one performance measure. The different performance measures are sometimes referred to as dimensions of performance; these terms will be used interchangeably in this report.

Performance measures can vary in several ways. First, they can be of differing complexity (e.g., a simple count of the number of defects on an inspection, or a supervisory rating of the leadership quality of a subordinate). Performance measures can also vary in terms of objectivity versus subjectivity. In the previous example, count of defects is fairly objective and easily quantified, whereas a rating of leadership quality involves more subjective processes and is less easily quantified. It is important not to confuse objectivity-subjectivity with the amount of judgment used to define the performance measure. A count of defects requires a considerable evaluative judgment before a clerk can make a tally. In this report, objective and subjective measures will be used within this definition, and no degree of judgment will be implied by either term. Subjective job performance measures are typically called performance ratings, and this convention will be followed. Objective performance measures are sometimes called production or productivity measures or records; however, that usage is somewhat erroneous. Production or productivity is much too general a term (as will be discussed later) to equate with the narrowness of objective performance measures. Further, it is obvious that subjective performance measures also indicate something important about an individual's productivity.

Finally, performance measures can vary in terms of the degree of control the individual has over altering personal performance on the measures. If there exist constraints on performance due to inadequate technology or supplies, for example, the individual can do little to affect performance on the measure. However, one would be hard-pressed to say an individual has similar constraints on a performance dimension such as personal appearance. Although there is no special terminology to differentiate performance measures on this continuum, this distinction will have considerable importance for establishing the foundation of the conceptual model.

The terms performance measurement, performance evaluation, and performance appraisal are used interchangeably in this report. They all refer to the process by which performance measures are generated. Kavanagh (1982b) has defined them as "the process, for a defined purpose, that involves the systematic measurement of individual differences in employees' performance on their jobs" (p. 192). This definition is consistent with other definitions found in the literature.

It is important to distinguish between a performance measure, performance measurement, and a performance measurement system. The performance measure is the outcome of the process defined as performance measurement. The performance measurement system involves all components of the performance measurement function within an organization. Thus, supervisory training to use the measures and the measurement process, administrative procedures for administering and maintaining the measurement, and the relationships between the performance appraisal system and other personnel systems are all included in this concept. Performance measurement system, performance evaluation system, and performance appraisal system are typically used interchangeably with performance evaluation program. In order to avoid confusion, only the term performance appraisal system will be used in this report to refer to the total function involved in measuring individual job performance.

The final point to be made here involves the use of the word productivity. Often, productivity measures are used interchangeably with objective performance measures. This is much too narrow, and productivity will not be used in this manner. Productivity is a more general term, and close to what Kavanagh (1982b) has defined as job performance. "Job performance is a dynamic, multidimensional construct, assumed to indicate an employee's behavior in executing the requirements of a given organizational role" (p. 195). The term job performance will be used in this report, following the meaning described above, to avoid confusion with other uses of the term productivity in the literature.

1.3 Background

Previous applied research efforts in the military have traditionally not used job performance as the dependent variable for validating personnel procedures and decisions. Typically, training school grades have been employed to validate the Armed Services Vocational Aptitude Battery (ASVAB) and its predecessor selection and classification tests. Although training school success is an important dependent variable in the military human resources management system, it is an intermediate criterion of the effectiveness of personnel selection. The crucial question that remains is whether the scores on the ASVAB can successfully predict individual effectiveness in job performance once the person is on the job. It is important to note that this logic applies not only to the validation of the ASVAB but to applied research efforts involving decisions within the human resources management system. For example, the empirical question of whether females can perform as well as males in traditionally non-female jobs in the military cannot be answered using training school success only. Nor can the effectiveness of a placement/transfer system be evaluated only in terms of personal adjustment and time to proficiency in the new job. Clearly, a measure of individual effectiveness on the job is needed to validate such personnel decisions. However, the focal example used throughout the remainder of this report will be the validation of the ASVAB.

If there exists a need for criterion measurement, why not use the performance appraisals already available in the military? The well-documented problems of leniency errors (or effects) and reduced variance for these ratings have limited their usefulness for validation efforts. More critically, the performance appraisals currently in use in the military are primarily designed for administrative actions (i.e., promotions). As such, they are subject to "gaming," which can distort the true score an individual should receive in terms of job performance. In order to evaluate the validity of the ASVAB, it is necessary to develop a performance measurement

methodology for research purposes only, so as to better estimate actual performance levels of individual airmen. The literature has shown that data gathered for research purposes, as opposed to data collected for administrative uses, contains less distortion and, more importantly, shows greater variance.

Thus, the need exists for a criterion measurement methodology to index individual effectiveness on military jobs for use in validation research. Over the past 2 years, technical reviews of the R&D programs of the AFHRL have consistently noted that this effort should not "start from scratch." The large volume of previous research on the "criterion problem," as well as the increased volume of available data as a result of the passage of the Civil Service Reform Act (CSRA) and recent Equal Employment Opportunity (EEO) court decisions, will serve as guidelines to enable the criterion development research to be accomplished in a relatively efficient and cost-effective manner. The purpose of this report, as noted earlier, is to provide a conceptual framework to guide this effort.

1.4 Classification Scheme Boundaries

To provide a clear focus for the classification scheme, it is necessary to describe the boundary conditions that will be used.

These conditions are:

1. The classification schema focuses on performance measurement in the military.

2. The schema describes the case in which performance measurement is being used for research purposes only.

3. The schema considers all variables that affect performance measurement: organizational, situational, group, dyadic, and individual.

In the military context, there may be performance variables that do not appear in non-military settings (e.g., those concerned with weapons maintenance and use, and with combat effectiveness). Also, some variables in the model may have greater salience than in the non-military context. In the military environment, too, the variance in performance ratings is likely to be greater than that found in non-military contexts. In non-military contexts, there is usually more pre-selection, and as a result, the variance on the selected aptitudes for jobs is smaller than that typically found in the military.

The fact that the performance measures will be used for validation research only will likely change the impact of the different variables (e.g., Zedeck & Cascio, 1982). In our schema, this means that the variables, their interrelationships, and their salience would change depending on the purpose of the measurement. In terms of a regression analogy, it would be expected that the beta weights for the independent variables would change as a function of whether the performance measures were to be used for employee growth and development, administrative, or research purposes. For example, the relation between pay and performance would be highly salient in a schema that was concerned with the use of measures for administrative purposes, but probably less salient within either a growth and development or validation framework.

The third boundary condition is not a constraint; rather, it is an extension of previous performance measurement schemas. Other classification schemes are typically more micro in their perspective. For example, some schemas, explicitly or implicitly, concern only the cognitive processes of the rater and their effect on the quality of the measures. Other approaches are concerned with the dyadic relationship between the rater and the ratee. The present

classification scheme will be more macro, and include all of the relevant variables that affect the measurement of individual job performance.

II. A CONCEPTUALLY BASED DESCRIPTIVE CLASSIFICATION SCHEME OF PERFORMANCE MEASUREMENT QUALITY

Based on the various considerations outlined in Section I, and an examination of the literature from the behavioral sciences, an approach to development of a classification scheme was selected that involves the following considerations:

1. Variables were included, on the basis of the theoretical and empirical literature, that could affect either job performance or the measurement of job performance.

2. Classical test score theory, with its emphasis on true and error variance in observed scores, provided a general perspective.

3. Rather than including detailed individual variables, these variables were classified into categories for ease of presentation.

4. An iterative process was used, beginning with a general schema of job performance and ending with a job performance measurement classification scheme for validation purposes.

5. The applicability of the classification scheme for use in a military setting was an overriding concern.

These considerations will be discussed as the schema is described.

Before describing the development process, we wish to emphasize that the frameworks for the schema, which were derived from the theoretical and empirical literature, are descriptive rather than prescriptive, because, in our judgment, the causal linkages hypothesized in the scheme are incomplete. This does not mean that no research evidence exists, but rather, that further research is necessary before the scheme can be classified as prescriptive. As will be seen in later sections of this report, there are a variety of research findings that impact directly and indirectly on the hypothesized linkages of the schema, and some of this literature can provide prescriptive advice for the development of a measurement methodology for job performance.

Perhaps the major reason for a conservative stance lies in the definition of measurement quality used in this report. As will be discussed in more detail, we consider accuracy and construct validity as the primary criteria for evaluating the quality of measurement when the purpose of the measurement is for validation research. Although other criteria of measurement quality (e.g., halo and leniency) have been used extensively to judge the "goodness" of job performance measures, we believe they have less relevance for "research only" performance measurement.

2.1 A Simplified Job Performance Schema

Prior to the development of a framework describing the measurement of job performance, it was necessary to develop a schema of job performance as a first step. It was important to identify all variables that could potentially impact on a person's job performance, since these same variables could be important sources of true or error variance in the measurement of job performance. An examination of theories in the area of work motivation (cf. Steers & Porter,

1979) showed them to be conceptually comprehensive, but lacking in detail in terms of the specific variables that affect job performance. However, using these general models and others from the organization behavior literature (e.g., Naylor, Pritchard, & Ilgen, 1980), a general schema of the variables that impact on individual job performance was constructed (Figure 1).

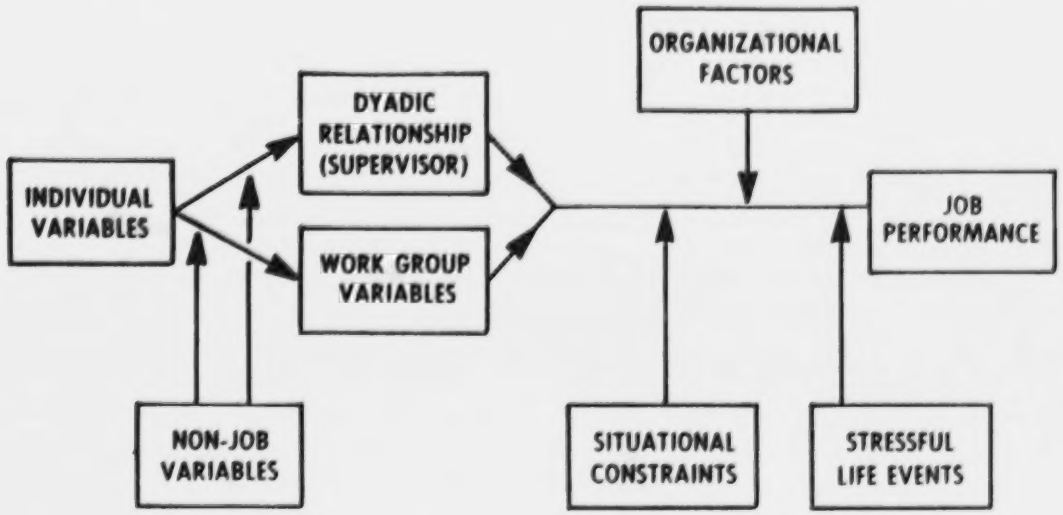

Figure 1. A Simplified Job Performance Scheme.

A brief description of the schema depicted in Figure 1 will suffice to provide the interested reader the opportunity to look more deeply into the theoretical underpinnings of the scheme.

Starting on the left side of Figure 1, the presence of individual variables in the model is axiomatic, and based on the common theme in the motivational literature (Steers & Porter, 1979) that individual job performance is a function of the skills, aptitudes, and effort a person brings to a job. These variables, according to Figure 1, indirectly influence job performance through their impact on the relationship with the supervisor (Bass, 1981; Vroom, 1976; Yukl, 1981) and their interaction with work group factors (Graen, 1976; Hackman, 1976). It is important to note that non-job variables could also affect the interaction of the individual variables with both the work group and the relationship with the supervisor. These factors would include such variables as marital status, religious preference, and membership in a dual-career family (Hamel, 1981; Owens & Champagne, 1965). Note that this class of variables is not on the major causal linkage in the schema, thus indicating that these variables may or may not impact at this point in the model. The same is true for the organizational factors (Adams, 1965; Lawler, 1976; Payne & Pugh, 1976), situational constraints (Chapanis, 1976; Peters, O'Connor, & Rudolf, 1980), and stressful life events (Kavanagh, 1982a). Although any of these factors can become quite salient in terms of affecting individual job performance, they are not always operative.

The "simplified" framework in Figure 1 is not only conceptually and empirically based, but it logically links major sources of variance in job performance together in a meaningful manner. Current research in the field continues to explore the importance of each of these variables; however, for our purposes, the differential impact of these variables on individual job performance is unimportant. As long as the possibility exists that each of the classes of variables in Figure 1 <u>can</u> influence individual job performance, then it is also possible that they can affect the quality of the measurement of job performance.

2.2 Development of a General Job Performance Measurement Classification Scheme

The model presented in Figure 2 provides a general classification scheme of performance measurement quality. The model in Figure 2 is similar to the Figure 1 model in that it suggests no direct, isomorphic relationship between a person's skills, aptitudes, and effort and the outcome variable. However, these input variables are included since they can contribute either true or error variance to performance measurement quality.

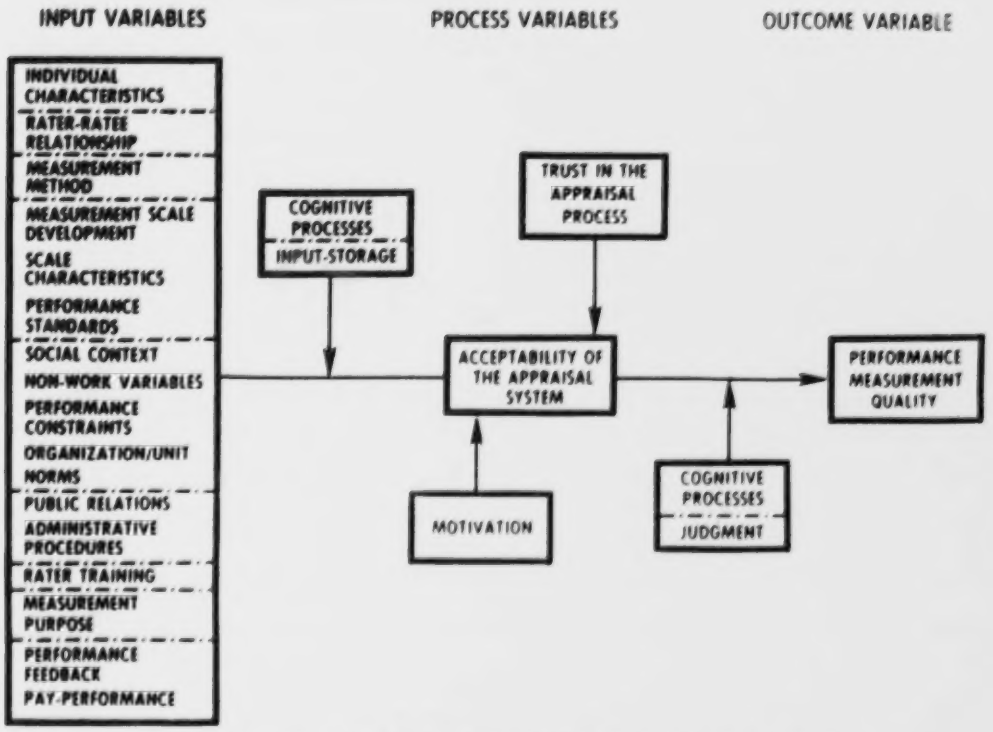

Figure 2. A Job Performance Measurement Classification Scheme.

It should be noted that Figure 2 is simply a general performance measurement quality classification scheme. A performance measurement framework for the purpose of validation research only will be discussed later. It is necessary to cover the more general case in order to understand the model-building process.

The general classification scheme in Figure 2 was developed by focusing only on those variables that impact the quality of performance measurement. First, six criteria that have been used to assess the quality of measures of job performance were identified. These are listed in Table 1. It should be noted that the first criterion is properly labeled psychometric "effects," not "errors," which we feel is consistent with current thinking in the field of performance measurement (Hakel, 1980; Hedge, 1982; Kavanagh, 1979).

Table 1. Quality of Performance Measurement Criteria

1.	Psychometric effects: halo, leniency, range restriction
2.	Inter-rater reliability
3.	Content validity
4.	Discriminability (in terms of individual performance levels)
5.	Construct validity
6.	Accuracy

Next, a literature search was conducted to identify the variables that impact these quality criteria. The variables identified constitute the input variables shown in the first box on the left side of Figure 2.

The process variables shown in the center of Figure 2 reflect the current thinking in the performance measurement literature that these variables play an important and pervasive role in the appraisal process (Borman, 1977; Dipboye & dePontbriand, 1981; Feldman, 1981; Hedge, 1982; Murphy, 1982). There has been a recent emphasis on cognitive variables (their importance in the decision-making process) (Feldman, 1981; Landy & Farr, 1980), as well as the acceptability/confidence users have in the system (Dipboye & dePontbriand, 1981; Kavanagh & Hedge, 1983; Landy, Barnes, & Murphy, 1978), and their hypothesized effects on measurement quality. In addition, the motivation which the ratees bring to the appraisal process (DeCotiis & Petit, 1978) and their trust in the appraisal process (Bernardin, Orban, & Carlyle, 1981) are considered important process variables. Although there is little empirical evidence in the literature with respect to the role of these variables, there are indications that they act as intervening process variables. Thus, although these individual/system characteristics are known to influence measurement quality, since they are hypothesized to be functionally related to both the independent and dependent variables, they will be considered separately.

The cognitive variables have been placed outside the main causal path since these variables may not always play an important role in the appraisal process. When the measurement system relies heavily on human judgment, such as with ratings or trained observers, these variables would be expected to influence measurement quality. However, when human judgment is not as important, such as with productivity counts or number of absences, the impact of these variables would be greatly reduced.

In Figure 2, the cognitive variables have been divided into two categories: (a) the input and storage of information, which is primarily concerned with the observational heuristics that people use when gathering information about an individual's job performance; and (b) the cognitive processes that involve the judgment or decision heuristics that people use in assigning a quantitative index to the performance of a person on the job. This division avoids the search for a single cognitive variable, such as cognitive complexity (Bernardin and Cardy, 1981; Lahey & Saal, 1981), that relates to measurement quality. Recent work by Murphy, Garcia, Kerkar, Martin, and Balzar (1982) and Hedge (1982) indicates that considering observational processes and decision processes separately may help to better explain their effects on the quality of the measurement. This is also consistent with Wherry's theory of rating (Wherry & Bartlett, 1982), which postulates that observation and recall by the rater are two separate components of the observed score.

A general hypothesis underlying this conceptualization is that the more complex (i.e., sophisticated, not necessarily cumbersome) the observational and/or decision heuristics used, the higher the quality of the performance measurement. However, individual/system characteristics could affect the complexity of these cognitive processes, and thus, lower or raise the quality of the measures. A good example is the impact of organizational or unit norms in the current military performance measurement system, where a strong norm exists to give enlisted personnel high ratings (i.e., "8" or a "9") on their performance evaluations. Regardless of the cause of this norm, its effect is to simplify the rater's cognitive approach; for whether or not the rater uses complex observational heuristics, the decision heuristic is simple -- "8" or "9." The impact on the quality of ratings is obvious, and interestingly, similar results have been found in many non-military settings where ratings are used for administrative purposes.

In the performance appraisal literature, few studies have focused on motivation in the context of performance measurement (Bernardin, Orban, & Carlyle, 1981; Bernardin & Cardy, 1982;

DeCotiis & Petit, 1978) or on trust in the appraisal process (Bernardin, Orban, & Carlyle, 1981). Still, the authors believe that these variables play key roles in the accuracy of performance evaluations; thus, both have been included as elements of the classification scheme.

User acceptance of and confidence in the performance measurement system are seen as crucial to the effective operation of the entire system, and thus, directly affecting the quality of the measurement (Kavanagh, 1982b; Lawler, 1967). Some recent empirical work (Dipboye & dePontbriand, 1981; Kavanagh & Hedge, 1983; Landy, Barnes-Farrell, & Cleveland, 1980; Landy, Barnes, & Murphy, 1978) indicates that this is an important variable in a performance measurement system. In our general conceptual framework of performance measurement quality, all of the system characteristics indirectly affect the quality of the measurement through their impact on the acceptability/confidence variable. Clearly, this acceptability variable may change in importance depending on the purposes of the performance measurement. This notion is critical to the development of a schema for validation purposes only, and will be discussed later in this section.

Another perspective used to generate the classification scheme was one borrowed from test score theory. Spearman's classic test score model was selected because of its simplicity and wide dissemination. The notion that an observed score, a performance measurement score, can be divided into true and error components allows us to examine the impact of those variables that affect true variance and those that affect only error variance in the performance measurement situation. During our literature review, it became obvious that one would want to minimize those factors that affect only error variance, while increasing the impact of those factors that influence true variance. This finding has clear implications for future research strategies.

This approach, based on test score theory, is analogous to that taken by Wherry (Wherry & Bartlett, 1982), although we prefer to base our conceptual framework on the Analysis of Variance model of test scores (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). In Wherry's theory, the observed rating a person receives is comprised of the following components: true job performance of the ratee, environmental influences, observation and recall by the rater, and the errors associated with these factors as well as an overall error term. Although the test score model used may not be critical, it can be seen from Figure 2 that the classification scheme is much more specific about the variables that impact on measurement quality.

Finally, in order to refine the classification scheme, our approach was to organize into categories the many variables that are known to affect measurement quality. This provided a framework for conducting the literature review in an organized fashion, and for identifying and prioritizing AFHRL research needs. This reasonably exhaustive list of variables is contained in Table 2.

As mentioned earlier, perhaps the most critical input variable in terms of its impact on rating quality is the measurement purpose. For example, if a measurement system is to be used for promotion or pay increases, it creates an entirely different context for the quality of the measurement than if the system is for validation research purposes. That is, the pay-performance relationship with measurement quality would be extremely important in a measurement system being used for administrative purposes, but would have little effect for validation research purposes.

Performance measurement systems have four major purposes or uses: (a) for administrative decisions, (b) for employee growth and development, (c) for validation research, (d) for meeting legal guidelines. The strength of the relationships between the individual/system characteristics and measurement quality will change as a function of changing the purpose. Although these effects have been discussed for some time (cf. Cummings & Schwab, 1973), only recently has there been empirical evidence which demonstrates that measurement quality is affected by the purpose of the measurement (Zedeck & Cascio, 1982). A good analogy would be to consider the individual/system characterisitics in Figure 2 as independent variables, the intervening variables as

Table 2. Variables That Can Impact on Measurement Quality

1. Individual characteristics
   a. Cognitive variables: rater or ratee
   b. Rater/ratee intelligence
   c. Rater/ratee knowledge of the job being evaluated
   d. Rater/ratee personal characteristics
   e. Rater/ratee interpersonal trust
2. Relationship between ratee and rater/observer
   a. Sex congruence
   b. Race congruence
   c. Job tenure together
   d. Age congruence
   e. Off-the-job relationship
   f. History of conflict or cooperation
3. Method/source of measurement
   a. Supervisor ratings
   b. Peer ratings
   c. Self ratings
   d. Subordinate ratings
   e. Assessment center (team) ratings
   f. Work samples/simulations
   g. Productivity records
4. Scale development
   a. Critical incidents used
   b. Based on job description/job requirements
   c. Employee participation
   d. Top management support during development
5. Rating scale characteristics
   a. Content of the scale
   b. Anchors versus no anchors
   c. Behaviors versus traits
   d. Format type
   e. Number of anchors/scale points
   f. Single versus multiple dimensions
   g. Scaling metric/approach
6. Performance standards/goals
   a. Present or not
   b. Standards versus goals
   c. Participately set and communicated
   d. Specificity of behavior or accomplishment expected
7. Social context
   a. Performance level of others in work group
   b. Existence of group norms
   c. Rater's status in group
   d. Ratee's status in group

Table 2. (Concluded)

8. Non-work variables
   a. Marital status
   b. Dependent Status
   c. Dual-career family
   d. Participation in company activities off the job
   e. Stressful life events in recent past
9. Performance constraints
   a. Poor information
   b. Equipment efficiency
   c. Supplies deficiency
   d. Time limitations
   e. Poor work environment
10. Organizational/unit norms
   a. Expectation of certain level of performance by upper management
   b. Expectation by immediate supervisor regarding level of performance
   c. Presence of a union
   d. Pay/rewards tied to performance levels by contract
   e. Pay/rewards tied to performance levels by informal norms
11. Public relations/administrative procedures
   a. Required or not
   b. Mode of presentation
   c. Content of procedures
12. Training
   a. Content of training
   b. Format of training
   c. Length of training
13. Measurement purpose
   a. Validation research only
   b. Employee growth and development
   c. Administrative purposes such as rewards
   d. To meet legal guidelines
14. Performance feedback
   a. Required or not
   b. Sources of feedback
   c. Participative
   d. Clarity of feedback
   e. Frequency of feedback
15. Pay-performance relationship
   a. Are they related in the system?
   b. Equity of the relationship

moderators, and measurement quality as the dependent variable in a multiple regression equation, and to expect the beta weights to change for the various terms in the equation as the purpose of the measurement changes. Since the initial thrust of the AFHRL program is for research validation only, we will now examine what happens to the general framework in Figure 2 when only this purpose is considered.

2.3 Validation Research Classification Scheme

The performance measurement classification scheme for validation research is depicted in Figure 3. It should be understood that the relationships among variables are the same as those described for Figure 2. However, there are some important differences between the two figures. First, the measurement purpose input variable drops out since Figure 3 is a validation research only classification scheme. Likewise, since research is the purpose, the performance feedback and pay-performance variables are omitted from Figure 3. Also, it should be noted that the acceptability of the appraisal system variable is now seen as less influential, based on the logic that when job performance data are collected for validation purposes only, user acceptance may not be as serious an issue. Thus, in Figure 3, while this variable is still central to the performance appraisal process, its impact on measurement quality is likely to be reduced. Although the variables and their relationships are comparable to those described earlier, there are several important aspects of Figure 3 that need to be discussed.

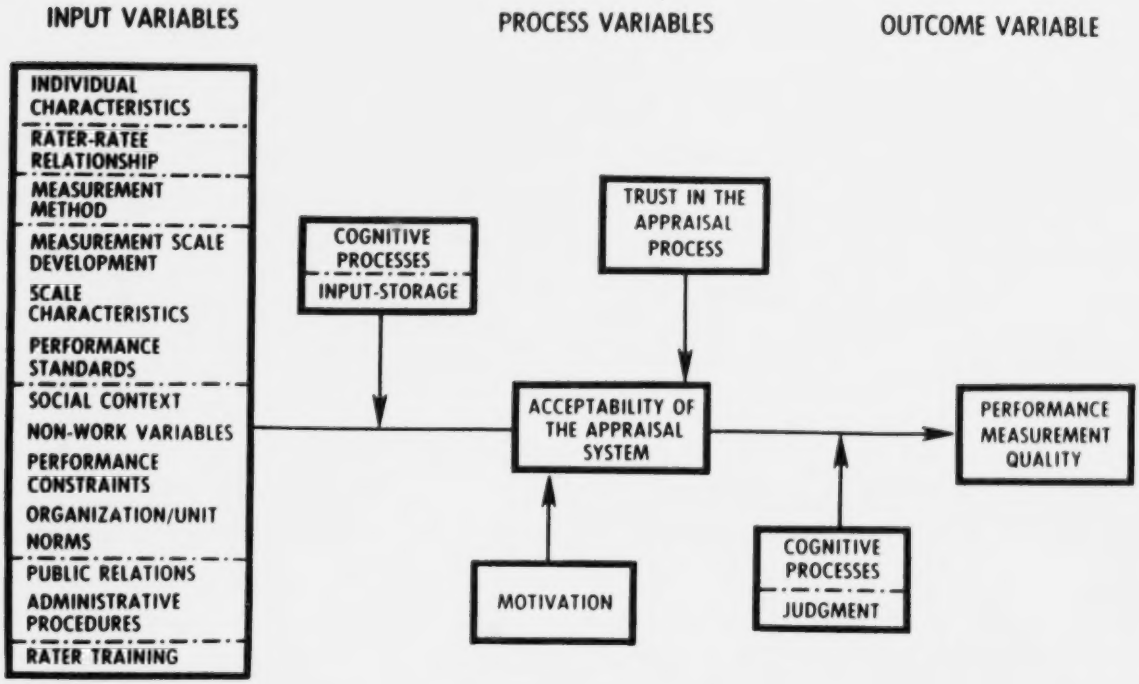

Figure 3. A Job Performance Measurement Classification Scheme for Validation Research.

First of all, the dependent variable, measurement quality, is something that has not been clearly defined in the literature. Different researchers have used differing criteria to assess measurement quality; six of these criteria were listed in Table 1. Of those listed, we view accuracy and construct validity as the crucial criteria by which to judge the quality of the measurement of job performance. The other four criteria are seen as important, but less critical, in that satisfying their requirements does not guarantee the measure will be accurate and construct valid. On the contrary, an accurate and construct valid measure will in all likelihood satisfy the other criteria as well. This logic is consistent with current theory in performance measurement (Nunnally, 1978) and performance ratings (Wherry & Bartlett, 1982).

Few models of the performance appraisal process exist in the literature. Of these, only two theoretical approaches were found that emphasized accuracy as the crucial criterion of measurement quality. One such theory was advanced by DeCotiis and Petit (1978), who argued that accuracy in ratings is a function of rater motivation, rater ability, and the availability of appropriate rating standards. Although most of the variables in their model also appear in Figure 3, their emphasis is on how the ratings are made, whereas ours is on the accuracy of the measure. Another difference is that the DeCotiis and Petit model concerns only ratings, whereas our conceptual framework encompasses any measure that is used to assess individual job performance.

Wherry and Bartlett (1982) also provided a model of the performance appraisal process, but as with the DeCotiis and Petit (1978) model, the model by Wherry and Bartlett is concerned only with ratings. Thus, the present schema is more comprehensive than either of the other two. However, as earlier noted, there is considerable similarity between the Wherry and Bartlett model and the present schema in terms of their bases in test theory.

Of the input variables shown in Figure 3, one that is critically important is the measurement method. As with the measurement purpose variable described earlier, constraints in terms of different methods will likely affect the relationships within that framework. If this is so, it may be that different measurement methods are capturing different parts of the performance criterion space. That is, supervisory ratings may well be assessing a different portion of the total job performance criterion space than are peer ratings, self ratings, work sample tests, or objective indices of productivity.

This is not meant to imply that there is no overlap among these methods in the part of the criterion space they measure; however, they are perhaps measuring some unique aspects of the criterion space that have been treated frequently in research as error. In the typical research paradigm to validate multiple measures of job performance, one or more methods have been eliminated because of low intercorrelations with the other methods, on the assumption that these low correlations were the result of error in the measures. In our schema, we conceptualize different measurement methods as measuring different parts of the criterion space with differing degrees of fidelity; thus, low correlations between measures may not indicate error. It can be argued that the typical validation approach (intercorrelations among methods) may not be the best for assessing measurement quality.

This discussion raises the issue of what, in terms of performance dimensions, constitutes the criterion space. Approximately 12 to 15 performance dimensions repeatedly appear in the literature. These dimensions seem to fit into two general categories: technical competence skills and job-relevant interpersonal skills. Although this may seem an oversimplification, it is supported by factor analytic studies in which two factors, roughly representing these two broad skill areas, have emerged (Borman, 1981; Borman, Mendel, Lammlein & Rosse, 1981). Although these two categories are, of course, multidimensional, viewing the criterion space in this manner provides an effective way of communicating AFHRL research needs.

In terms of measuring job performance, the following five methods are the most frequently used: (a) supervisory ratings, (b) peer ratings, (c) self ratings, (d) work samples, and (e) objective indices of productivity. The first three are widely used and will be used in this research effort. However, rather than using a traditional work sample methodology, an alternative to this approach will be developed and tested.

The new methodology is called Walk-Through Performance Testing (WTPT); it is being developed specifically for the R&D program at AFHRL. The WTPT methodology combines aspects of both observer interviewing and work sampling but, in addition, is designed to overcome certain limitations associated with the generic tasks used with work sampling. The method will be

developed by accessing the Air Force data base (see Christal, 1974) that contains information on the tasks performed in enlisted specialties. These tasks will form the basic content of the measurement scale. Test administrators will be trained to use these scales to evaluate effective and ineffective performance on each of the tasks. The interviewers will examine the job incumbent by asking the person to perform certain tasks or explain certain procedures concerning the tasks for that job. They will then record the person's behavior or answers on a rating checklist of tasks. The important characteristic of this method is that the job is being reduced to its smallest parts at the task level, and will include not only a core set of tasks, but a series of unique tasks as well. Thus, this method will examine job performance at micro level.

It is believed that the WTPT method will assess, with a high degree of fidelity, technical skills and competence -- one-half of the criterion space. In fact, walk-through testing may be one method that removes the interpersonal/social aspects of the job situation. However, as currently planned, it may be less accurate in assessing the job-relevant interpersonal skills side of the criterion space. Supervisory ratings, on the other hand, may be quite good at assessing interpersonal skills but not very accurate in measuring technical skills, particularly if the job has had recent significant changes in technology, or if the supervisor has never had direct work experience. All five methods measure portions of the criterion space; however, they differ in their fidelity or accuracy of measurement.

This does not mean that one or more of the methods could not be modified to assess both major parts of the job performance criterion space. The WTPT method could be modified, for example, to measure interpersonal skills. However, this modification may not be cost effective if there is another method that can accurately assess the interpersonal skills without modification (e.g., peer ratings). In summary, the five methods, as currently used, assess different parts of the criterion space with differing degrees of accuracy. AFHRL's criterion development research must be designed with this point in mind.

This approach to job performance measurement makes the typical multimethod validation study problematic. Typically, if job performance is measured with two or more methods, and zero-order correlations are calculated among the methods, those methods showing nonsignificant values are rejected. However, if, as has been argued, the methods are not assessing the same portions of the criterion space with equal fidelity, then there is no reason to expect them to be correlated.

An extension of this logic leads directly to the idea of specifying the construct space for job performance in terms of what Cronbach and Meehl (1955) have termed a nomological network (a network of relations that are tied to observables and hence are empirically testable). In this framework the measures are the observables, and the construct is used to account for relationships among them. This suggests the use of Nunnally's (1978) technique for construct validation; namely, that of testing the a priori hypothesized relationships within a construct space with empirical data. To do this, the two major parts of the criterion space, technical competence and interpersonal skills, must be better specified in terms of their job performance dimensions.

Having delineated a multidimension-multimethod matrix, the next step in this research strategy would be to hypothesize the expected level of relationship between each method-dimension and all others. In this manner, one would have specified, a priori, the hypothesized nomological net for these methods and the criterion space. After collecting data, the results would then be examined to verify the expected correlations. In this strategy, a zero relationship would be as important as a non-zero one in establishing the construct validity of the methods of measurement.

For the Air Force R&D program, empirical construct validation cannot be accomplished until at least the fifth or sixth year of the effort. First, the methods must be properly researched and

refined. There is a tremendous amount of research to be done during the first 4 or 5 years of the program before this type of study can be conducted.

2.4 Additional Considerations for the AFHRL Program

Among the research issues to be investigated prior to construct validation are two major considerations. The first of these concerns the WTPT methodology for assessing job performance. The second concerns the incorporation into the schema of an additional variable, time to proficiency.

As described earlier, WTPT is a combination of work sampling, simulation, and interviewing, which assesses job performance at the micro-task level. When fully developed, it is expected to have the highest fidelity for assessing some aspects of job performance. Thus, it can serve as a standard against which to judge the appropriateness of other methods for evaluating technical competence.

A research strategy underlying the AFHRL program might be referred to as successive approximations to high-fidelity measures of job performance. As currently envisioned, the WTPT method should accurately assess the technical job skills of individuals. However, this method is quite time-consuming and expensive, particularly for large-scale data collection across the Armed Forces. If research proves that WTPT does indeed have high fidelity for technical skills, we can then determine which of the less expensive and time-consuming methods of data collection on job performance most closely approximate the WTPT, and can be used instead. Also, as earlier noted, if the WTPT method can be modified to accurately measure interpersonal skills, the same research strategy of successive approximation with less costly and time-consuming methods can be used to tap this portion of the job criterion space.

Secondly, an additional variable, time to proficiency on job tasks, needs to be incorporated into the R&D program. Research is necessary to determine how to best adapt the five methods to measure this crucial part of job performance. A wide range of individual differences on this variable likely exists among newly assigned personnel, particularly in their first job in the military. Furthermore, it is likely that one or more of the methods can measure this variable across task performance with greater degrees of fidelity/accuracy. In fact, there probably is a task/dimension by method effect on the accuracy of assessing time to proficiency. The important point is that this variable must be considered in any research effort.

In this chapter, the development of a conceptually based classification scheme of performance measurement quality for validation research was described. This scheme, depicted in Figure 3, will be used to summarize and organize previous research as well as to specify needed future research. If one draws a line between any of the variables (either input or intervening process variables) and the dependent variable, a linkage in the schema is defined. In Chapter III, the literature will be categorized according to the appropriate linkage in the schema, and conclusions regarding the known empirical "facts" within each linkage will be drawn. This will allow a specification of what research is still needed for effective criterion development in validation work (Chapter IV).

III. LITERATURE REVIEW

3.1 Introduction

As previously mentioned, an extensive literature search was conducted, resulting in a voluminous collection of citations and abstracts. The review was concerned only with literature on the assessment of the quality of job performance measurement systems.

The focus of the review was on behavioral literature, preferably empirical research, concerned with investigating the hypothesized linkages in the schema in Figure 3. This was accomplished to identify well-documented "facts," which could form the basis of prescriptive advice for the AFHRL R&D program. That is, when we prescribe that a given system characteristic must be included in the measurement methodology, there is no need for additional research by the AFHRL to "re-establish" this finding. On the other hand, the literature review also identified where research is needed for the AFHRL program.

Although the review was limited to literature on measurement system quality within the performance appraisal field, the entire scope of available writings was considered. In some cases, the search was simplified by identifying previously published reviews (Kane & Lawler, 1979; Kavanagh, 1971; Landy & Farr, 1980; Lewin & Zwany, 1976; Mabe & West, 1982; Schmidt & Kaplan, 1971; Smith 1976). Some of the literature covered by these reviews was not relevant to the purposes of this report, but where it was, the review was used as a secondary source.

3.2 Individual Characteristics and Measurement Quality

This linkage concerns the differences between raters or observers (in the WTPT method) that can impact on the accuracy of the measures of individual job performance.

Eighteen studies were identified that examined relationships between rater/observer characteristics and measurement quality. Two studies (Borman, 1977; Mullins & Force, 1962) found consistency in rating accuracy across two different jobs and two different job performance dimensions, respectively. Murphy, Garcia, Kerkar, Martin, and Balzar (1982) found that accuracy in observing behaviors is related to accuracy in evaluating performance; however, they noted that this relationship may be more complex than their results showed. This study will be examined more closely when rater cognitive variables are discussed. It does appear, nonetheless, that there are important individual differences in raters that affect their rating accuracy.

The major study investigating the relationship between individual differences and rating accuracy was done by Borman (1979a). Borman found that 12 personal characteristics correlated significantly with rating accuracy. The most consistently high correlations were between accuracy and (a) intelligence, (b) personal adjustment, and (c) detail orientation. It is important to note that the variance in accuracy accounted for by all 12 traits was 17%, suggesting that individual differences play a significant role in determining rating accuracy. In a meta-analysis of self-evaluation studies, Mabe and West (1982) found that intelligence, achievement motivation, and internal focus of control were associated with accuracy in self-evaluations. Other research has found that rating quality is related to: carefulness and decisiveness of the rater (Mullins, Seidling, Wilbourn, & Earles, 1979), learned associations among sets of behavioral and personality descriptors (Hakel, 1974), and interpersonal trust (Kavanagh, Vance, & Wright, 1982). Clearly, the individual traits and characteristics of a rater are related to measurement quality.

In a slightly different vein, two recent studies investigated individual differences derived from laboratory data versus those derived from field data. In a study of police supervisors, Kavanagh et al. (1982) found no statistically significant relationships between accuracy, leniency, halo, and range restriction based on ratings of videotapes with true scores (Borman, 1979b) gathered in the laboratory and leniency, halo, and range restriction of ratings used operationally in the field. Hedge and Kavanagh (1983), using the same approach, found only one relationship, halo from laboratory and field data, to be significant. These unexpected findings need further replication, and it may be that other individual characteristics of raters are moderating these relationships.

Recent literature (Feldman, 1981; Landy & Farr, 1980; Wherry & Bartlett, 1982) has called attention to the cognitive processes involved in evaluating others. Using halo, leniency, range restriction, and rater confidence, Schneier (1977) found that cognitively complex raters were more confident in making their ratings and also made fewer leniency and range restriction errors. Using the same dependent variables, two subsequent studies (Lahey & Saal, 1981; Sauser & Pond, 1981) found no support for a relationship between cognitive complexity and measurement quality. Finally, in three further studies using both the typical psychometric errors and accuracy as the measurement quality criteria, Bernardin, Cardy, and Carlyle (1982) found no support for a hypothesized relationship with cognitive complexity. It seems clear that cognitive complexity, as measured in these studies, is not related to measurement quality.

Does this mean that the cognitive processes of the rater do not affect the quality of the measurement? In a simplistic approach using a single trait such as cognitive complexity, the answer is probably "yes." However, raters use multiple cognitive processes (Landy & Farr, 1980; Wherry & Bartlett, 1982) in arriving at a specific judgment, and the literature supports studying these processes separately (as indicated in Figure 3) in terms of their effects on measurement quality. Hedge (1982) found that training raters in either observational techniques or decision-making techniques differentially affected psychometric errors and rating accuracy. Murphy, Martin, and Garcia (1982) found that ratings made 1 day after viewing videotapes with true scores were more accurate than ratings made immediately following the showing of tapes, and suggested that different memory processes were in operation. Related to our model in Figure 3, it may be that the immediate ratings involve only observational heuristics, whereas the delayed ratings involve decision heuristics.

Evidence that there are individual differences in the decision processes of raters comes from several studies. Policy-capturing studies have identified raters who were consistent in their rating policies (Hobson, Mendel, & Gibson, 1981), raters who used linear regression models in their strategies (Zedeck & Kafry, 1977), and raters who used different decision strategies depending on the purpose of the measurement (Zedeck & Cascio, 1982). The latter study, although supportive of our model, unfortunately did not include validation research as one of its purposes of measurement. Finally, one study found that different raters (self, peer, and supervisor) used different aspects of performance in arriving at their evaluations of performance (Zammuto, London, & Rowland, 1982). Taken together, these studies indicate that there are important individual differences in both observational and decision processes that can affect the quality of measurement. Future research should treat these two sets of cognitive variables separately when attempting to determine their effects on the quality of performance measures.

3.3 Rater-Ratee Relationship

Much of the research reviewed in this and the previous section also applies to observers in the proposed WTPT method. Within the performance appraisal context, most of the research has been done on raters; however, the processes that WTPT observers go through are quite similar in that they also require observation and judgment. They differ in that observers have more specific, job-relevant events to observe, and judgments occur immediately following the observation of the events; whereas raters must make judgments based on job events recalled from some previous time period. Nevertheless, it is probable that many of the findings for raters will generalize to observers. Where they do not, or when special considerations must be made for the role of the observer, it will be noted.

Research on the rater/ratee-rating quality linkage concerns the formal and informal relationships between raters and ratees, as well as the influence of the sex, race, and age of the ratee on the quality of the measure. Although most research has concentrated on the interaction of the rater's sex, race, and age with the sex, race, and age of the ratee, some

research documents the main effects of these variables. In the opinion of the authors, there is always an interaction between raters and ratees on these biographical variables; however, not all of the relevant research investigated the interactions.

Several studies examined characteristics of the relationship between raters and ratees. The degree of responsibility the rater had over the ratee's previous performance (Bazerman, Beekun, & Schoorman, 1982), the rater's familiarity with the ratee's previous performance (Jackson & Zedeck, 1982; Scott & Hamner, 1975), and the degree of acquaintance between rater and ratee (Freeberg, 1968) have all been shown to affect the quality of job performance measurement. The degree of acquaintance variable is perhaps most interesting. It is axiomatic that the rater must be at least somewhat knowledgeable about ratee performance to complete the rating. In fact, most authors argue that the rater must have had the opportunity to observe job-relevant behaviors or else the rating will contain error (e.g., see Borman, 1975). However, Stone (1970) has argued that as the degree of aquaintance increases, the possibility of bias in terms of halo increases, particularly if the rater and the ratee become friends. This logic is consistent with Wherry's theory of rating (Wherry & Bartlett, 1982); however, it has not been directly tested in the performance measurement domain.

Three studies (Freeberg, 1968; Gordon, 1972; Quinn, 1969) have partially addressed the issue of rater-ratee acquaintance; however, none of these studies included the full range possible for the acquaintance variable. Freeberg (1968) could not find any effect due to length of acquaintance, and concluded that degree of acquaintance is only important for rating quality when it provides greater opportunity to observe job-relevant behaviors. Quinn (1969) found no effect for acquaintance, but the range of scores on length of acquaintance was not sufficiently large to adequately test the hypothesis. Neither of these two studies included friendship as a variable that could impact on rating quality. Gordon (1972), in a study tangential to this issue, found that the favorability of the rater's impression of the ratee did not affect the accuracy of the ratings. Favorability of impression may be close to friendship; unfortunately, Gordon's study was a laboratory simulation using college students as the subjects, thus limiting the generalizability of the results. It seems clear that well-designed studies are needed to explore the acquaintance/friendship variable.

Many studies have investigated the effects of differences in the sex, race, or age of both raters and ratees on the quality of performance measures. In such research, it is assumed that these biographical variables are causing error in the measurement, but few studies have used accuracy as the criterion as defined in this report. Evidence of bias, rather, comes from level differences that indicate certain groups are receiving lower ratings on the basis of their demographical characteristics.

Only two studies examined the influence of age on performance measurement quality (Cleveland & Landy, 1981; Schwab & Heneman, 1978). For ratings of "paper people" vignettes of four secretaries varying in age, Schwab and Heneman (1978) found no age-associated differences in the accuracy of ratings. Cleveland and Landy (1981), examining ratings from 513 exempt managers, found that older workers were rated lower on two performance dimensions, self-development and interpersonal skills. They replicated these findings on a second sample. It is not clear, however, whether Cleveland and Landy's results indicate differences in accuracy for ratings of older workers, or actual differences in the performance of the older workers in the sample. Finally, it should be noted that these two studies did not examine the interaction with rater age, and its effect on measurement quality.

Research results related to sex effects on performance measurement quality have been inconsistent. Two studies (Bigoness, 1976; Hamner, Kim, Baird, & Bigoness, 1974) found that females were rated higher than males, and this effect was accentuated when they were performing at high (versus low) levels. However, both were laboratory studies using videotapes and

employing students as raters. In two field studies (Cascio & Phillips, 1979; Mobley, 1982), sex was found to have little effect on performance measures. Although females were rated higher than males in the latter study, neither study found any effect of sex differences on personnel decisions. In another laboratory study, Lee and Alvares (1977) found no differences in ratings as a function of sex. Feild and Holley (1977), in a field study, found that sex differences existed within specific job classes, but one could not generalize these differences across all job classes. Consistent with this finding, Nieva and Gutek (1980), in a review of the empirical literature, concluded that evaluation bias by sex does exist, but its effects are not consistent across all situations. Their more specific conclusions indicated that the degree of bias is a function of the level of inference required in the rating, sex-role incongruency with the job, and level of qualification or performance. Of particular importance for this review are the effects of sex-role incongruency on the accuracy of performance measures collected for validation purposes. The level of qualifications or of performance in a job can only serve to increase sex-role incongruency; thus, it should also be included in research on this possible bias. The level of inference required can be controlled by the scale development, and that will be discussed later in this section.

Results on race are also somewhat inconsistent. Various studies have found significant bias effects due to race (Bigoness, 1976; Hall & Hall, 1976; Hamner et al., 1974; Feldman & Hilterman, 1977; London & Poplawski, 1976). Other studies have not found these race effects (Bass & Turner, 1973; Cascio & Phillips, 1979; Cascio & Valenzi, 1978; Farr, O'Leary, & Bartlett, 1971; Greenhaus & Gavin, 1972; Huck & Bray, 1976; Mobley, 1982; Moses & Boehm, 1975). It is interesting to note that these latter studies were done in field settings, whereas the former were done in the laboratory. Wendelken and Inn (1981) noted this distinction, and conducted a major field study investigating race main effects and the interaction between raters and ratees. They found significant effects for ratee race, rater race, rater-ratee interaction, and past performance of the ratee; however, these four effects accounted for only 4% of the variance in the ratings. As a result, they suggested that the larger effects found in the laboratory are artifacts. It appears that race of rater or ratee, or the interaction, may well affect the measurement quality of ratings collected for validation purposes; but, the effect is expected to be small. To the extent that this variable may impact on the accuracy of the measures, it needs to be investigated.

The impact of other variables involved in the rater-ratee relationship on measurement quality have also been demonstrated. Drory and Ben-Porat (1980) found that leadership style, specifically initiating structure, was related to leniency on task-related dimensions in rating subordinates. Cascio and Valenzi (1977) found that rater-ratee experience and education both impacted significantly on ratings, but they also concluded that the practical significance of the results were not very important. Beatty, Schneier, and Beatty (1977) found that ratees perceived desired job behaviors as occurring more often and undesired behaviors occurring less often than did raters. Finally, Barrett (1964) found that supervisors give high ratings to subordinates who do their work the way the supervisor wants it done, regardless of what the job standards indicate. All of these studies indicate the importance of the rater-ratee relationship in determining the quality of ratings; however, no single finding appears important enough to merit further individual study when performance measures are collected for validation purposes only.

3.4 Measurement Method

It is assumed that all of the measurement methods that are currently available in the behavioral science literature suffer from criterion deficiency, some being worse than others (e.g., production records or counts are probably worst). Further, it is posited that each of these methods measures a part of the criterion space with more accuracy than other measures. For example, peer ratings may be the most accurate method for assessing interpersonal skills that are job relevant; yet when used to assess other parts of the criterion space, often all that is

measured is error variance. That is, the various measurement methods have frequently been used to assess parts of the criterion space which they are ill suited to measure. This leads to two conclusions. First, in research using multiple methods to assess individual job performance, the empirical demonstration of a zero relationship is as important as a non-zero one in demonstrating the construct validity of the measures. The second conclusion is that one must use multiple methods of measurement to accurately assess the total criterion space.

The industrial/organizational psychology literature tangentially supports these arguments. Research comparing multiple sources (Baird, 1977; Basset & Meyer, 1968; Blackburn & Clark, 1975; Borman, 1974; Griffiths, 1975; Holzback, 1978; Ilgen, Peterson, Martin, & Boeschen, 1981; Kavanagh, MacKinney, & Wolins, 1971; Klimosk & London, 1974; Kraut, 1975; Meyer, 1980; Schneier & Beatty, 1978; Thorton, 1980; Wiley & Hahn, 1977; Zammuto et al., 1982) has found (a) disagreement among factor structures for different rating sources, (b) differences in rating strategies, and (c) low discriminant validities for the measurement methods. On the basis of his results, Borman (1974) argued that a "hybrid" rating system should be created in which raters make evaluations on only those dimensions they are in a good position to rate. Likewise, Schneier (1977), in reviewing the literature, concluded that it is erroneous to collapse ratings across raters, and that the use of multiple sources could improve rating accuracy. Thus, there is support for the argument that a measurement methodology for assessing individual job performance should include multiple sources, with each measurement source measuring only that part of the criterion space for which it has the highest fidelity.

There has been no research attempting to develop a multiple-method approach to the assessment of job performance. As a first step, it will be necessary to determine, for each measurement method, which part of the criterion space it can best measure. Though all measurement methods, including production records, may be assessing some true variance in the criterion space, the problem is the present uncertainty about which methods tap which parts of the criterion domain; this should be the first research conducted. There has been good research done on the various methods, and some hypotheses can be developed about their use in a multiple-method appraisal system. Once each of the methods can be refined sufficiently and it can be determined that they are assessing certain parts of the criterion space with a high degree of accuracy, the next step should be to investigate the feasibility of developing and using a multiple-method approach for the measurement of individual job performance.

Earlier, we listed five methods of performance assessment. It should be noted that there are other methods for the assessment of job performance; however, these are the methods that have been used most successfully in the past.

The evidence supporting supervisory ratings is considerable. In addition to the multiple-source studies cited earlier, Landy and Farr (1980) provided a comprehensive review indicating the acceptability of supervisory ratings. Supervisors are often in the best position to assess the person's overall contribution to the effectiveness of the work unit. That is, the supervisor may be best qualified to weigh the person's performance across the various parts of the criterion space to reach an overall judgment.

Support for the use of peer ratings is also available in the literature (cf. Downey, Medland, & Yates, 1976; Fiske & Cox, 1960; Kaufman & Johnson, 1974; Lewin & Zwany, 1976). Peer ratings have had their best success at predicting future behavior, and it is hypothesized that this is so because they accurately assess those job-relevant interpersonal "survival" skills that are important for successful performance across jobs.

In addition to the multiple-method studies that have included self ratings of performance, evidence supporting the usefulness of self ratings appeared in a recent review (Mabe & West, 1982). These authors concluded that, controlling for measurement conditions (e.g., specificity

of the measurement instrument, amount of prior self evaluation experience), self ratings can provide good measures of abilities.

The fourth method, WTPT, combines interviewing and observing, and includes elements of work sampling, simulation, and work observation techniques. There is strong support in the literature for the use of all of these techniques for the measurement of individual job performance (Boehm, 1982; Hakel, 1982; Robertson & Downs, 1979; Smith, 1976; Vineberg & Joyner, 1982). As discussed previously, this method will focus on measuring the smallest possible unit of job-relevant behaviors. The test administrator will not only be involved in the observation of specific job tasks, but will also ask the job holder to describe verbally how he or she would deal with a specific job-related task or problem. Thus, the method will assess job-related skills, and also perhaps interpersonal skills, and even supervisory skills. It is hypothesized that this method, given proper development, will provide the most accurate measurement of the technical competence portion of the criterion space.

The fifth method, production and other objective records, is included for a variety of reasons. First, such measures are usually readily available and are commonly used as indices of the effectiveness of units and organizations. There is support for their use in the literature (Ronan & Prien, 1971), particularly in jobs that require production quality and quantity counts. But most importantly, they can be used to capture an important part of the criterion space. Typically, these types of measures indicate the employee's compliance with certain work or organizational rules. Violations of these rules, though infrequent, are what creates the measurement problem. However, aggregating these records in some form may provide important information on individual work performance that is not being measured with the other methods. Whether these individual differences in work behavior represent a lack of compliance or other motivational problems remains to be determined, but it is hypothesized that measuring them will improve the overall assessment of the criterion space.

The research directions seem clear. If the measurement methodology used for validation research is to assess job performance with minimal criterion deficiency and maximum accuracy, then research to develop and validate each of these methods is necessary. Success in these efforts will lead to a multiple-method research effort concerned with establishing the differential accuracy of each measure for portions of the criterion space.

3.5 Performance Standards, Scale Characteristics, and Scale Development

With the exception of the production records, all of the methods require scales to measure job performance. Since three linkages in the classification scheme in Figure 3 are concerned with the measurement scale (i.e., performance standards, scale characteristics, and scale development), the evidence for all three will be reviewed in this section.

Based on their review of legal cases regarding compliance with Equal Employment Opportunity Commission (EEOC) guidelines concerning performance appraisal forms, Cascio & Bernardin (1981) argued that the performance appraisal form must have performance standards if it is to be in compliance with the guidelines. Because this recommendation is for operational systems, it may not apply to performance measurement used only for validation purposes. However, if the existence of performance standards can improve the accuracy of the measurement, then they should be used regardless of the purpose of the measurement. There are arguments for the use of performance standards (Alewine, 1982; Kirby, 1981; Morano, 1979) but unfortunately, no empirical results that would support their use. Performance standards on the rating format may make the form easier to use by raters/observers, and thus enhance its accuracy; however, this has not been tested empirically, and remains an avenue for research.

As a number of authors have noted, the search for the single best scale format for measuring job performance or the best type of content for the scale has resulted in no conclusive evidence (Kavanagh, 1971, 1982a; Kingstrom & Bass, 1981; Landy & Farr, 1980; Muczyk & Gable, 1981; Schwab, Heneman, & DeCotiis, 1975). This does not mean that a researcher or a practitioner can be cavalier about the selection of a format or other scale characteristics when developing a performance appraisal scale. In fact, as all of the cited reviewers have noted, care in development of the measurement scales is more important for the quality of the measure than is the specific format or content chosen. The one lesson that the enormous literature on scale development and scale characteristics has demonstrated is that there are right and wrong ways to develop performance measurement scales. This section will focus on the literature supporting these prescriptions for scale development, noting where research may be needed when the purpose of measurement is for validation only.

In terms of scale development, it seems clear that the scales must be based, in some way, on job descriptions (Cascio & Bernardin, 1981) and job task requirements (Cornelius, Hakel, & Sackett, 1979; Rosinger et al., 1982; Tosti, 1979). It also seems clear that the participation and support of management (Beer, Ruh, Dawson, McCaa, & Kavanagh, 1978) improves the quality of the developmental process and thus, the quality of the measurement. Although it is quite common to include both raters and job incumbents in the scale development process, the evidence supporting this approach is meager and indirect (Friedman & Cornelius, 1976; Williams & Seiler, 1973). Most of the benefits derived from rater and employee participation are a result of their increased acceptance of the system. For validation purposes, acceptance of the system is probably crucial if one expects to obtain accurate measures. Thus, it appears that both employee and rater participation in scale development may be necessary.

Regarding scale content in general, there seems to be no clear resolution as to whether one should use personal traits or performance dimensions only (Kavanagh, 1971; Massey, Mullins, & Earles, 1978). However, when job performance is the target domain to be measured, the only conceptually appropriate content is performance dimensions. As Kavanagh (1982b) noted in a review of some of the early research contrasting behavioral anchored rating scales (BARS) with non-anchored Graphic Rating Scales (e.g., Borman & Dunnette, 1975; Burnaska & Hollman, 1974; Campbell, Dunnette, Arvey, & Hellervik, 1973; Keaveny & McGann, 1975; Maas, 1965), the critical feature of the rating scale is how clearly the performance dimensions are described.

In the multitude of studies reviewed concerning development of alternative forms for the measurement of job performance (Arvey & Hoyle, 1974; Beatty et al., 1977; Bernardin, 1977; Bernardin, Alvares, & Cranny, 1976; Bernardin & Smith, 1981; DeCotiis, 1977; Dickinson & Tice, 1977; Fay & Latham, 1982; Finley, Osburn, Dubin, & Jeanneret, 1977; Githens & Elster, 1973; Ivancevich, 1980; King, Hunter, & Schmidt, 1980; Latham & Wexley, 1977; Mullins, Weeks, & Wilbourn, 1978; Nugent, Laabs, & Panell, 1982; Rizzo & Frank, 1977; Rosinger et al., 1982; Saal & Landy, 1977; Schwartz, 1977; Seaton, 1974; Shapira & Shirom, 1980; Siegel, 1982; Zedeck, Kafry, & Jacobs, 1976), several rather strong conclusions emerged. First, the anchors or descriptors that define performance levels on job dimensions must be observable job behaviors or accomplishments. Second, these observables must be related to job-relevant tasks. Third, the scale must be structured such that the rater can use it easily. Fourth, the format used is not as important as these other characteristics. Finally, if an overall measure of job effectiveness is to be collected, it should occur at the end of the form, after individual dimensions of job performance have been assessed.

Some experts (Kavanagh, 1980, 1982b; McAfee & Green, 1977) have argued that in selecting an appraisal format, a broader class of criteria should be used in addition to the traditional psychometric ones. McAfee & Green (1977) have contended that the various performance appraisal formats available are differentially useful, depending on the purpose of the measurement. For validation purposes, they suggested that direct indexes (objective or administrative measures

such as AWOL rate), weighted checklists, forced choice, or other variations on rating scales can be appropriate. Kavanagh (1980, 1982b) listed 19 different criteria that could be important in choosing a scaling format for the measurement of job performance. It seems important that these additional criteria be carefully considered in the validation effort.

Also, at least one expert has shown that going beyond five scalar points does not increase the reliability of the measures (Smith, 1976). However, there is also evidence (Finn, 1972) that up to seven scalar points improves the quality of the measures in performance ratings. It seems that the number of scalar points is not an important research issue as long as the scale contains at least five. However, the decision to have more scalar points should not be based on psychometric properties alone. There may be other important reasons such as improving rater acceptance of the form.

Two other important research issues remaining to be resolved concern scale content and the scale development process itself.

The content issue involves such questions as: Are 10 to 12 performance dimensions adequate to cover the job performance criterion space? Are there performance dimensions that are common to all jobs within an Air Force Specialty (AFS)? There is some evidence that there may not be universal dimensions for job performance (Feild & Holley, 1975), and that different raters may define the job performance criterion space differently (Borman, 1974; Taylor & Wilsted, 1974; Zedeck, Imparato, Drausz, & Oleno, 1974).

Based upon the literature, and the authors' experience in building performance measurement systems, several hypotheses have been formed concerning this content issue. First, many jobs contain two general categories of performance dimensions, technical skills and job-relevant personal and interpersonal skills. Second, there are a number of universal job dimensions that are common to all enlisted jobs; for example, combat readiness and communications skills. However, the way in which persons in different jobs perform their job-relevant tasks may be quite different. To take a simple example, the communication skills required for effective job performance as an aircrew mechanic would be decidedly different from those for a crew chief or a clerical job. Third, the number of job dimensions required to cover adequately the job performance criterion space is probably not more than 12 for non-supervisory positions and 15 for supervisory positions. Obviously, the speculations voiced here need to be empirically verified.

Finally, a number of authors have raised questions about the techniques typically used to develop rating scales (Bernardin & Kane, 1980; Dickinson & Tice, 1977; Kane & Bernardin, 1982; Kavanagh & Duffy, 1978; Latham, Saari, & Fay, 1980; Schwab et al., 1975; Shapira & Shirom, 1980). The development of behavior-based rating scales starts with the generation of "critical incidents" of job-relevant behaviors and accomplishments. These are usually collected from job incumbents; however, they could be collected from supervisors, job knowledge experts, or peer employees in parallel but different positions. When a performance measurement system is to be used for validation purposes, a crucial research question is: Which of these sources of critical incidents will lead to development of the most accurate measurement scale. As a second step, behavior-based rating scale development typically involves a second set of judgments or a statistical clustering to determine which are the best critical incidents and job performance dimensions to put into the final rating form. Again, the research issue is which of these various techniques will result in the most accurate performance measure for use in validation research. These are important research issues for the rating and WTPT methods being considered for AFHRL.

3.6 Environmental Context, Non-Work Variables, Performance Constraints, and Organization/Unit Norms

These four classes of variables, as identified in Figure 3, will be considered here as a single category since their combined effect generally results in increased error variance. These variables are similar to the environmental influences term in Wherry's theory of rating (Wherry & Bartlett, 1982); however, our approach has greater specificity, allowing for better identification of research issues.

There is ample evidence that environmental and situational factors can affect the quality of performance measures (Borman, 1978; McCall & DeVries, 1976; Schwab et al., 1975; Scott & Hamner, 1975; Turnage & Muchinsky, 1982). There is also evidence of the impact on measurement quality of specific situational variables such as unit norms (Grey & Kipnis, 1976), situational constraints (Peters, Fisher, & O'Connor, 1982; Peters & O'Connor, 1980; Peters, O'Connor & Rudolf, 1980), and social context (Knowlton & Mitchell, 1980; Mitchell & Liden, 1982; Wood & Mitchell, 1981). Among the non-work variables, there is documented evidence that marital happiness and environmental stressors such as duration on task directly impact job performance (e.g., Rose, Jenkins, & Hurst, 1978; Sharit & Salvendy, 1982; Wilkinson, 1969).

In terms of non-work variables, environmental stressors, and performance constraints, it is important to determine if these are contributing to error variance in the measures. Or, alternately, are raters adjusting their judgments because they are aware of the existence of these factors and their temporary negative effect on the person's true level of performance? For example, does a supervisor adjust his/her ratings of the job performance of an employee because the employee has just had a death in the family? It may be that instructions or training emphasizing such factors may be sufficient to minimize their potential contribution to error variance in the measures. This will be discussed more fully in a later section; however, it is apparent that these factors must be considered in developing an accurate measurement system.

Some research must also be conducted to ensure that the final measurement system is not being adversely affected by social context or the existence of unit or organizational norms. Because of the importance of organizational and unit norms in the military, these variables should receive particular attention.

3.7 Public Relations and Administrative Procedures

The notion that confusing administrative procedures and instructions to raters will diminish the quality of the measures can be inferred from the literature. It seems apparent that the clarity of presentation and the administrative procedures for scales with observable job-relevant behaviors has led to the relative success of this type of format. Two major efforts involving the development, implementation, and evaluation of performance measurement systems (Beer et al., 1978; Kavanagh, DeBiasi, Hedge, & Miller, 1983) provide indirect evidence of the importance of these variables. In both efforts, the importance of a public relations program to "pre-sell" the performance measurement system was emphasized. They also noted the importance of forms management, clarity of instructions to raters, and ensuring the administrative procedures for the measurement system were consistent with other personnel procedures in the organization. Although the literature generally has paid little, if any, attention to these variables, it is believed they are crucially important to the acceptance and accuracy of a performance measurement system for validation purposes. Because the AFHRL system must, at some point, be used for large-scale data collection, the importance of these variables is amplified.

Finally, as noted previously, it may be possible to minimize error variance by the use of instructions or public relations. For example, a public relations program which emphasizes that

the performance measurement system is for validation research only, and that raters should disregard the organizational norms that predispose raters to judge everyone an "8" or a "9," may be effective in minimizing this potential source of error. Likewise, clear instructions to raters to not let off-the-job factors influence their evaluations of employees may be sufficient to control for this potential error source. It seems clear that this type of research will be necessary if the performance measurement system is to generate accurate ratings.

3.8 Rater Training

This is one of the most important linkages in the classification scheme in Figure 3. Though most of the research reviewed deals with rater training, the principles are equally applicable to observer training in the WTPT method. However, observer training is not seen to have as many research issues as rater training since observer training is much longer, more under the control of the researcher, and can thus do a better job of eliminating problems that negatively affect the quality of ratings.

Although rater training does not always significantly impact measurement quality, most empirical studies show positive effects of training on both the traditional psychometric concerns and accuracy (Bernardin, 1978; Bernardin & Pence, 1980; Bernardin & Walter, 1977; Borman, 1975, 1979b; Brown, 1968; Fay & Latham, 1982; Hedge, 1982; Ivancevich, 1979; Latham, Wexley, & Pursell, 1975; Levine & Butler, 1952; Sauser & Pond, 1981; Spool, 1978; Stockford & Bissell, 1949; Taylor & Hastman, 1956; Thornton & Zorich, 1980; Warmke & Billings, 1979; Zedeck & Cascio, 1982). Clearly, it is not necessary to conduct research to determine if rater training should be a part of a performance measurement system; rather, the important research issues involve the specific characteristics of the training.

Length of the training is a research issue. Although there is little evidence that strictly applies to the accuracy criterion, the literature indicates that training sessions as short as 5 minutes can improve the quality of the ratings (Borman, 1975). Zedeck and Cascio (1982) trained students over 5 contact hours using the typical "psychometric error" approach, and found no effect on accuracy using the "paper people" approach. Hedge (1982), using real raters, found significant effects on accuracy with a 2-hour training session. Thus, the issue of length of training has not been resolved, and may be complicated by the type of training given.

In those studies that used the traditional psychometric effects as the criteria of measurement quality, it appears that participative techniques (group discussion, videotaping, role playing) are better than lecture only. Training raters to maintain diaries appears to also improve measurement quality (Bernardin & Walter, 1977). In those studies that used accuracy as the criterion of measurement quality, it appears that the traditional "psychometric error" training approach did not have a positive effect (Hedge, 1982; Zedeck & Cascio, 1982).

However, training raters to be better observers (Hedge, 1982; Thornton & Zorich, 1980) or better decision makers (Hedge, 1982) has been shown to have positive effects on the accuracy of the measures. Though not extensive, the evidence certainly suggests an important research issue for rater training.

A research issue that has received no attention is who should do the training. For large-scale data collection in the Air Force, it may be impractical to use personnel psychologists to train all raters. It may be possible to use senior enlisted personnel as trainers at their respective bases with no loss in terms of the quality and effectiveness of the training. This issue will obviously have critical implications for AFHRL's validation research effort.

A final issue in rater training (and observer training) is possible "wash-out" effects of training. Even with lengthy and intensive training sessions, raters return to making the same errors they did before training (Ivancevich, 1979; Latham et al., 1975). There has been no research on "booster" or refresher training of raters or observers, and this would also seem an important research question.

3.9 Intervening Process Variables

Of the five intervening variables shown in Figure 3, the two cognitive variables were discussed earlier; thus, the discussion in this section will focus on the acceptability, trust, and rater motivation issues. Research has shown that the acceptability of performance appraisals and appraisal systems can significantly impact measurement quality (Dipboye & dePontbriand, 1981; Kavanagh & Hedge, 1983; Landy et al., 1978). When a performance measurement system is being used for validation research, user confidence that accurate judgments can be made about job performance may affect measurement quality. This issue needs to be resolved through empirical research.

Rater motivation has been essentially ignored by performance appraisal researchers. This may be the result of a general belief that individual differences do not exist across raters in their motivation to rate accurately. Although DeCotiis and Petit (1978) incorporated rater motivation as an important part of their model of the appraisal process, they cited only Taft's (1971) theory of interpersonal judgments as support for the inclusion of this variable in the model. Recently, Bernardin and his colleagues (Bernardin & Cardy, 1982; Bernardin, Orban, & Carlyle, 1981) have focused on rater motivation, but only in terms of how it might be affected by the level of trust a rater has in the appraisal system.

It is hypothesized that there are differences in motivation and trust in the appraisal system across raters. In addition, it seems logical that appraisal accuracy may be affected by the interaction of motivation or trust and such system characteristics as purpose of appraisal, administrative procedures used, and appraisal format. For example, rater motivation to provide accurate ratings may be higher when ratings are gathered for research purposes rather than for administrative or developmental purposes. In addition, for a "research purposes only" system, rater motivation to rate accurately may depend on the administrative procedures/public relations used in implementing the system. These and other similar issues need to be resolved empirically.

IV. RESEARCH IMPLICATIONS

As discussed in Chapter I, the major objective of the classification scheme and literature review was to provide guidelines for specifying research needs in developing a performance measurement system for validation research in the military. Implications for research will be of two major types: (a) specification of areas where research is not needed; that is, where the literature provides prescriptive advice on system characteristics; and (b) specification of areas where further research is needed; that is, where the empirical evidence is inconclusive.

We have taken a conservative approach to accepting that a given system characteristic will impact on accuracy, especially when evidence relates only to the impact of the other rating quality indicators. On the other hand, where there are a number of studies examining a particular system characteristic, and the results are consistent across studies, it will be concluded that no further research is needed, even though accuracy was not used to assess measurement quality. These recommendations will be based on the judgment that the empirical evidence is strong enough to warrant generalizing to the accuracy criterion, and further, that

the cost of the research would be a waste of resources. In any case, we have recommended that research be conducted in those important areas where we believe there to be inadequate empirical evidence.

4.1 Individual Characteristics

It seems clear that rater/observer individual differences will affect measurement quality. Some variables shown to be related to measurement quality are the personal characteristics of the raters (Borman, 1979b), accuracy of the rater as an observer (Murphy, 1982), carefulness and decisiveness (Mullins et al., 1979), learned associations among sets of behavioral and personality descriptors (Hakel, 1974), and interpersonal trust (Kavanagh, Vance, & Wright, 1982). Although cognitive complexity, as commonly measured, is not related to measurement quality (Bernardin & Cardy, 1981), this does not mean that cognitive processes are not; dividing them into observational and decision processess, as done in the conceptual model, has some support in the literature. The literature strongly supports the importance of cognitive processes in performance measurement; and the work of Hedge (1982) on decision processes and accuracy in ratings, as well as Murphy's (1982) work on observational processes and accuracy, indicates these are avenues to pursue.

Not all of the research directions suggested by the literature are relevant for the development of a measurement methodology for job performance, particularly one to be used only for validation purposes. For example, Borman's (1980) suggestion that we should select raters on individual differences that are related to rating quality is an intriguing research project, but one that may not be feasible or beneficial for the Air Force research project.

A major research effort should be concerned with the relationship between the cognitive processes (heuristics) involved in observing and judging. Research in this area should be sequential. The first step would be to determine if different observation and decision processes are differentially related to accuracy of performance ratings and/or observations. The next step is to determine if rater/observer training can effectively teach raters/observers to use these processes. This research effort could be quite important to the WTPT method. The work of Hedge (1982) would seem a good starting point to explore both processes and their relationship to accuracy.

Some of the earlier cited research on individual differences might be useful in selecting WTPT observers. Evidence indicates that some individual differences do affect the accuracy of both observations and judgments about human performance. To ensure high fidelity in the WTPT method, it will be necessary to be sure that the instrument that records the data (the observer) is a source of only minimal error. Thus, it might be quite appropriate to select observers. However, selecting raters may be more difficult.

The study by Murphy (1982) relating observational accuracy to rating (judgment) accuracy, although limited in terms of the sample used, has some interesting implications for research. If, within the military setting using performance measurement for research purposes only, it can be established that observational accuracy is strongly related to rating accuracy, then it may be possible to use observation tests to evaluate how accurate raters are likely to be in assessing performance.

Although evidence points to a number of individual differences that are important in determining measurement quality, it would seem fruitless to investigate variables that are not easily changed via training. This does not hold true for the selection of WTPT observers. Observers may be selected according to scores on individual differences variables found to be correlated with accuracy. In sum, individual characteristics may not be important for rating

methods (unless training can modify them); however, they may form the basis of selection of observers in the WPTP method.

Other variables that are important to the accuracy of both the rater and the observer are their knowledge of the job and the degree of acquaintance with the job incumbent. Rater/observer understanding of the job is clearly important for rating quality. The degree of acquaintance between rater and ratee will be discussed in the next section.

4.2 Rater-Ratee Relationships

Even when performance measurement is for research purposes only, rater-ratee relationships would be expected to impact the quality of the measures.

From the literature review, it is apparent that the rater's familiarity with the ratee's previous performance (Jackson & Zedeck, 1982; Scott & Hamner, 1975), as well as the rater's acquaintance with the ratee (Freeberg, 1968), will impact the quality of the measurement. This would seem to be one of the most important areas of needed research for both the WTPT method and the peer and supervisory rating methods. For the WTPT process, it means that observers should probably be selected who are not knowledgeable about ratees' previous performance. Degree of acquaintance is an area needing research. It appears that observers need to become acquainted with the ratee's job performance in terms of relevant behaviors and/or accomplishments. However, as the literature indicates, too much acquaintance with the ratee can lead to bias in the ratings (typically halo). The crucial research question for WTPT observers is how much behavior they need to observe to make an accurate judgment about job performance. There would appear to be an optimum point (or a range) in terms of the amount of information about job performance required to make accurate evaluations. Too little information will lead to errors of deficiency; too much information may lead to biases.

From the literature, it is known that sex, race, and age can impact on the quality of judgmental data (e.g., Bigoness, 1976; Hamner et al., 1974; Nieva & Gutek, 1980). However, not all laboratory or field studies found these effects (Cascio & Phillips, 1979); in field studies such as those of Mobley (1982) and Wendelken & Inn (1981), only about 4% of the variance was due to these effects. Whether the same effects will occur when the measurement is made for validation purposes is unknown, but this is clearly a research issue. Possible ways to minimize these effects are through training and through effective scale construction that focuses on relevant performance behaviors. The latter approach seems an appropriate research area for the Air Force.

For the WTPT method, it will be necessary to determine if there are sex, race, or age effects on the quality of the evaluations. Further, the extent of any observer-ratee interaction on these variables should be determined. Finally, for both the ratings and the WTPT method, it will be necessary to determine if there are sex effects on measurement quality when there is sex role incongruity, particularly for females in non-traditional job specialties.

4.3 Measurement Method

In this proposed program of research, recall that the focus is on the following measurement methods of job performance: (a) supervisory ratings, (b) peer ratings, (c) self ratings, (d) WTPT, and (e) objective indices of productivity. The first three are well known. WTPT, a new method developed specifically for this project, combines both observer interviewing and hands-on methodologies. Test items will be constructed to evaluate performance on the required tasks for enlisted specialties. Test administrators, trained to use these scales, will examine job

incumbents by asking them to perform these tasks, or answer specific content questions concerning the tasks. The job incumbent's behavior or answers will then be recorded on a rating checklist. Thus, this method will examine job performance at a micro level.

As discussed earlier, it is believed that the performance criterion space for many jobs consists of two major parts--technical competence and job-relevant interpersonal skills. Both are necessary for effective job performance. However, no single measurement method is able to accurately assess both parts equally well. In fact, it is hypothesized that the different measurement methods assess different parts of the criterion space with varying degrees of accuracy.

The first, and perhaps, most critical research task is to specify these hypothesized relationships among the measurement methods and the dimensions of job performance within the criterion space. This will involve the generation of an "expected" correlation matrix within the multitrait-multimethod scheme. In line with various recommendations in the scientific literature (e.g., Feldman, 1981; Nunnally, 1978) that a priori conceptualizations must drive research programs, this is seen as a high priority task early in the research program; however, empirical testing of this nomological net will not occur until much later in the research program. Obviously, the measurement methods must be developed through research efforts such that there is reasonable certainty that the methods are accurately assessing the portion of the criterion space for which they are intended.

This research task would involve an intensive examination of the literature on measurement methods and specific dimensions of performance. Although we covered much of the literature in the present effort, our purpose was simply to identify those measurement methods that have been used with success. A more intensive literature search would be needed to form hypotheses about what specific portions of the criterion space are being accurately measured by the specific methods.

The completion of a conceptual framework of interrelations among method-dimension measures requires that a related research issue be investigated simultaneously. This issue involves determining the content of the criterion space (e.g., How many job dimensions constitute this space, and is there agreement on a general set of performance dimensions that is common across jobs within the military? Further, are there additional job dimensions that are common within some job families but not common within other job families?)

The literature and past experience suggest there are a set of common job dimensions that apply to most jobs, and that successful performance on these common dimensions probably changes across job families and job levels. For example, the job-relevant social skills for an aircraft mechanic would be expected to differ from those for clerical specialties and for supervisory jobs. Although communications skills are required in each, the job behaviors and performance standards would be different for the three jobs. Kavanagh et al. (1983) found this pattern of similar dimensions of performance (with differing content within dimensions) across job families in developing a performance measurement system for a multi-hospital corporation. It is reasonable to expect to find such a pattern in the development of the performance measurement methodology for the Air Force. This research effort will be concerned with reviewing the literature to identify an initial set of common job dimensions that can be used to generate a nomological net of relationships among method-dimension measures.

It has been shown that each of the five methods being considered within this research program accurately assesses some portion of the criterion space. Thus, none of these methods should be abandoned early in the research program; rather, research should be aimed at determining how to refine these methods so that they achieve a high degree of fidelity in measuring their respective portions of the criterion space. It is important that these methods be evaluated within an

accuracy paradigm, whether that be via multitrait-multimethod, "paper people," or videotapes with true scores. The traditional measures of measurement quality should also be collected; however, all research within this program must use at least one accuracy measure to evaluate the quality of measurement. This is not only important conceptually; it is crucial if one wants to generalize results across the various research projects. The establishment of specific, acceptable research paradigms using an accuracy criterion is critical to the success of this entire program of research.

Research on objective indices of individual job performance should proceed somewhat differently than research on the other methods. Objective indices in this case refers to items such as administrative counts (e.g., court martials, absences, inspection records, and accident rates). There are several problems with these indices, however. They have low base rates, suffer from criterion deficiency, are unreliable indices of true events, and typically, show low variance. This does not mean these indices should be ignored, because they may be capturing a piece of the criterion space; however, no major research effort on these measures is warranted. Perhaps when the other methods have been refined, these objective indices will become part of a larger data collection effort and prove useful for comparison purposes. Finally, by closely coordinating with an Army Research Institute (ARI) effort that is currently examining these objective indices, AFHRL may save considerable time and effort spent on evaluating objective measures.

A good organizational plan would be to treat the five measurement methods as separate streams of research, each with decision points to "stop," "modify," or "continue." Of course, there will be a point in the research program where AFHRL will begin to integrate these methods into a measurement methodology for job performance. Thus, one could envision these streams of research as the rows in a matrix describing the total research program, with the columns representing research activities or tasks within the streams. One could then establish decision points within each stream as to the quality of the method.

4.4 Measurement Scale Development

The following prescriptive guidance exists for developing rating scales:

1. Both raters and ratees should be involved in the development of the scale content.

2. The rating scales must be based on job descriptions/requirements.

3. A "critical incidents" approach should be used in initially identifying the relevant observable behaviors and accomplishments that define differing levels of job performance. The special strength of this technique is the high degree of content validity in the resultant scales.

4. It is crucial to have visible, top management support for the research program.

4.5 Scale Characteristics

This is the linkage in the classification scheme about which the most empirical research and fairly clear prescriptive advice exist, as follows:

1. The dimensions of performance should be well defined, and anchored by observable behaviors or accomplishments.

2. There should be at least five scale points for the rater/observer to make a judgment. Although reliability does not increase much beyond five scale points, having a larger number of points may increase raters' acceptability of the scale. It would probably be wise to avoid a nine-point scale because of its similarity to the current operational measurement system. Thus, a five- or seven-point scale should be used. And, it is important that scales across the methods use the same number of scalar points.

3. Each scalar point needs to be anchored with observables.

4. The type of format used is not as crucial as the other characteristics described above.

5. An overall measure of effectiveness should be collected, probably at the end of the form, after the individual dimensions of performance have been assessed. This should be done with the rating methods and the WTPT method.

In terms of needed research, the content issue raised in the "Measurement Methods" section should be a priority. Identification of common dimensions across jobs and a general mapping of the criterion space for Air Force AFSs will be important. Research, in coordination with an intensive literature search to identify common performance dimensions, should help to better identify the needed content for the measurement scales. Other work within AFHRL, such as research on skill, difficulty, and aptitude requirements for various AFSs may be applicable for identifying the performance dimensions needed to adequately measure the criteron space. Coordination of these research efforts should remain an AFHRL responsibility.

4.6 Performance Standards

Performance standards are an important part of this measurement methodology. A performance standard establishes a specific level, in terms of observables, at which a job incumbent must perform to be classified "acceptable," "unsatisfactory," etc. Performance standards also have important implications for scale development. It may be that we do not need performance standards, since we are measuring for validation purposes only; the important thing here is to accurately rank-order job incumbents according to their performance levels. Performance standards are not needed to accomplish this. On the other hand, use of standards might increase user acceptance of the scales. This is clearly an unresolved issue which needs some attention in this research program.

For the walk-through testing method, the research issue is how scales are to be developed such that observers can make objective judgments about a person's level of performance on job tasks. The issue becomes one of generating levels of performance on specific tasks that can be observed and coded in an objective manner that will define differing levels of performance. This appears to be close to the performance standards issue raised here, and the resolution may be research aimed at determining which group(s) of persons (job knowledge experts, etc.) are best qualified to provide input for development of the observer scales and performance standards.

4.7 Social Context

From the limited research reviewed in this linkage of the schema (Grey & Kipnis, 1976; Knowlton & Mitchell, 1980; Mitchell & Liden, 1982; Wood & Mitchell, 1981), it appears that social context contributes to error variance. Since none of the studies involved validation research, it may be that the effects of social context will not occur, or will be greatly minimized, in a validation effort. Using the laboratory paradigm discussed earlier in relation to age, race, and sex effects, the effects of social context could also be investigated. This would seem to be an important research project for the Air Force program. If these effects are present in a

research-only context, then research must determine how to eliminate these errors when data are collected in the field. Again, this may be accomplished by training, or more simply, by instructions.

The importance of these social context effects may be most relevant to the peer rating method, both with respect to the contrast effect and the biasing of job skill ratings by the ratee's level of interpersonal skills. The contrast effect occurs when a ratee's true performance level is altered because those with whom he/she works have performance levels different from his/hers. For example, an "average" performer may be rated higher if other persons in his/her work group are low performers. This has been found for supervisory ratings, and may be even more powerful for peer ratings. Clearly, this is a research issue that needs investigation. Another biasing effect is when an individual's performance evaluation in the technical skills area is altered by the rater's appraisal of his/her interpersonal skills. Peer ratings might be particularly subject to this bias, and research needs to be done to investigate whether these effects will occur in a validation research situation.

Interpersonal skills may also impact the evaluation of technical skills in the WTPT method. This possibility could be easily addressed in a laboratory study. By creating an experimental situation where the ratees are approximately equal in job skills but quite different in their interpersonal dealings with the observer, it would be possible to assess the impact of interpersonal skills on the judgments of technical competence. If the WTPT method is to have the highest fidelity of all methods for assessing technical skills, then there should be research to determine if the interpersonal interactions between the job incumbent and the observer are affecting the observer's judgments. If this is so, then research will be needed to help eliminate these errors.

4.8 Non-Work Variables

Non-work variables include marital status, religion, pre-school children in the family, a working spouse, and stress events. Based on the job-related stress literature, it seems clear that these variables do impact on job performance, but it is not at all clear that they influence ratings of job performance. These variables tend to be taken into account by supervisors in completing their ratings. That is, when there are non-work events that have a temporary effect on an individual's job performance, raters usually adjust for these factors. One way to minimize non-work factors is to instruct the raters to consider the person's performance over a longer period of time. Past experience indicates that this is likely to "factor out" the contaminating influences of non-work variables; however, this still needs empirical testing. Also, this may be a more serious problem with the WTPT method. If, for example, a person is having a problem at home and this has temporarily reduced his or her job performance, the observer will not know to adjust for that factor. It may be necessary for the observer to collect information from the supervisor on these possible effects before collecting job performance data.

4.9 Performance Constraints

As the literature indicates, performance constraints can also affect the individual performance of job incumbents. As with non-work variables, raters typically take these constraints (machines, supplies, forms, etc.) into consideration when evaluating an individual. These factors pose the most serious threat to the WTPT method. Thus, any research on this method should take these variables into consideration.

4.10 Organization/Unit Norms

This variable has a particularly large impact on rating quality when the ratings are used for administrative purposes. Though this may not be an issue if the measures are collected for research validation only, it will be crucial that the purpose of the ratings be clearly and strongly communicated to the raters. This may become troublesome for large-scale data collection. It may mean the use of a public relations campaign and strongly worded instructions to the raters that the data are being collected for research purposes only. This may be not so much a research issue as a practical issue.

The issue of effectively communicating the purpose also impacts on the packaging of the performance rating materials. These materials need to be packaged and presented in such a way that the purpose and procedures are completely understandable to the raters. Frequently error is introduced in performance appraisals simply because raters do not understand the instructions. There is little evidence on this issue in the scientific literature; however, it is believed that it can be a major source of error in the performance ratings.

4.11 Public Relations/Administrative Procedures

Although there is little research in these areas, it has been noted several times in this report that these factors play an important role when large-scale data collection is being conducted for research purposes. This should be an area of continuing concern throughout the developmental phases of the various methods, particularly so later in the research program.

4.12 Rater Training

The literature rather consistently indicates that some type of rater training is needed to ensure high quality ratings. This training can be quite elaborate, involving videotaping and experiential learning, or it can be more simple, consisting essentially of instructions to avoid certain rating errors. In the present context, it may be possible to provide orientation training to senior enlisted personnel in the proper use of the rating form. It may be that accurate measures can be obtained with relatively simplistic training. This is clearly a research issue, and one that needs to be investigated using accuracy as the primary dependent variable. Obviously, for large-scale data collection, using on-site trainers with relatively short orientation sessions would be the most desirable from a cost-effectiveness standpoint.

The work by Hedge (1982) indicates that if more extensive training is necessary to increase accuracy, alternative training approaches to traditional "psychometric error" training must be investigated. This effort should come as a second step in the research program. For example, the following rater treatments might be tested: (a) no training, instructions only; (b) orientation training and instructions; (c) psychometric error training, orientation, and instructions; (d) observation training, orientation, and instructions; and (e) decision-making, orientation, and instructions. Although this is a rough design, it emphasizes the point that research must be conducted to determine which of these or other procedures can best improve the accuracy of the ratings for validation purposes. This research can use either "paper people" or videotapes with "true" scores in a laboratory setting for the initial work.

As discussed previously, any factors identified as contributing to error variance in the judgmental data must be reduced through some countering technique. Training is one such technique, as are instructions to raters, packaging of the forms, and public relations efforts. It will require some careful scientific judgment, based on available research results, to decide

which of these techniques will best counter the error effects of these factors. Where possible, research studies should be conducted to help make these decisions.

Two other important training research issues need to be investigated: duration of training and "wash-out" effects. In terms of the length of the training session, there is little in the literature to suggest a minimum amount of time necessary to effect substantial improvement in accuracy. This probably depends to a great extent on the content of the training. Training sessions as short as 5 minutes can improve the quality of ratings but are not as effective as 1-hour training sessions. At least one study has shown that accuracy can be improved with a 2-hour training session. Obviously, this issue is far from settled, and it is a critical one for large-scale data collection. The length of training is a direct cost item. Tangentially, in terms of cost, research should also be conducted to determine if on-site enlisted personnel can be used as trainers, rather than using professional trainers. Though this may mean the creation of training manuals and sessions to train the trainers, the cost savings over using professional trainers are enormous.

The literature documenting wash-out effects has shown that even with intensive and lengthy training sessions, raters return to making the same errors they did prior to training. There has been no research on "booster" or refresher training, and this would seem to be an important research issue. This research could easily be a follow-on study to the training research study described earlier. The experimental subjects would simply be randomly assigned to either a refresher or no-refresher training condition after 6 months. Results would indicate the usefulness of refresher training for improving rating quality.

Training for the WTPT observers should be based on the sources of error variance identified in previous research in this program. Training will probably need to be intensive to bring observers to a high level of accuracy in their observations and judgments.

4.13 Intervening Variables

In the classification scheme depicted in Figure 3 are five process variables which may impact on the quality of the measurement: (a) observation heuristics, involving the input and storage of performance information; (b) decision heuristics, involving weighing information and making a judgment about a level of performance; (c) rater motivation; (d) rater trust in the appraisal process; and (e) acceptability of the measurement system.

Within the performance measurement literature, research on observational processes (Hedge, 1982; Murphy, 1982), decision processes (Borman, 1977; Hedge, 1982), cognitive processes (Feldman, 1981), rater motivation and trust (Bernardin, Orban, & Carlyle, 1981; Bernardin & Cardy, 1982), and acceptability of the rating system (Dipboye & dePontbriand, 1981; Kavanagh & Hedge, In press; Landy, Barnes, & Murphy, 1978) indicates these can be important determinants of rating quality. All of this research was conducted either in a laboratory or a field setting using operational measures of job performance. The extent to which these variables will operate in a system used for research purposes only is unknown at this time. It would appear that laboratory research would be a first step to determine if these variables can affect the accuracy of the measures. It would then be critical to collect field data in follow-up research. These variables are seen as high priority items, particularly if it is possible through training to alter these processes and significantly improve the accuracy of the measures.

4.14 Measurement and Research Paradigm Issues

Several important methodological issues which seemed to be quite relevant but did not fit neatly into any of the linkages of the schema remain to be discussed.

The first issue concerns the ways in which the accuracy of the measures of job performance are assessed. Though the other criteria for evaluating the quality of the measure (see Table 1) will typically be collected in all research projects, the crucial issue revolves around the selection of an accuracy approach. The four approaches to the accuracy/construct validity criterion are: (a) the multitrait-multimethod analysis (Kavanagh et al., 1971); (b) videotapes of performance, with known true scores (Borman, Hough, & Dunnette, 1976); (c) "paper people," with known true scores (Zedeck & Cascio, 1982); and (d) specification of expected score distributions on an *a priori* basis, as discussed in the introduction of this report. Two obvious questions are: (a) Are they the same; that is, will the same conclusions be reached about the accuracy of the measure regardless of the accuracy approach used? and (b) Should research be conducted within a laboratory setting to evaluate each of these?

A second issue concerns the "paper people" and videotape approaches. Can the materials created by other investigators be used within the military context to assess the accuracy of ratings? If not, then research must be started to create new videotapes and/or "paper people" that are specific to the military. For example, it may be necessary to create videotapes of aircraft mechanics engaged in job behaviors that vary in terms of true scores. Or, it may be necessary to create videotapes and/or "paper people" depicting military supervisors (or incumbents in other jobs).

The third issue is closely allied to the second. If the decision is made to use videotapes or "paper people" created by other investigators, will it be necessary to re-establish true score profiles using military rater-experts? Do the true scores for these materials generalize across organizations, particularly when there are important differences between military and non-military settings? Obviously, if materials are created that are specific to the military only, true scores will be generated by military rating experts, presumably subject-matter experts or supervisors. If these materials are created for military jobs, they could also be used to evaluate and train the observers in the WTPT method.

No definitive answers to these questions/concerns were found in the literature reviewed. Some research will be necessary to assess the adequacy of videotape versus the "paper people" approach. It is the opinion of the authors that in order to serve as viable tools, these videotapes must be Air Force specific.

The last issue is a most critical one. Careful control must be exercised over the research conducted in this program. If specific approaches to the accuracy criterion are used, they need to be used in all research studies. This would also hold true if other criteria are used to evaluate the quality of the measures of job performance. Consistency is critical if the results of these separate research projects are to be combined to arrive at operational decisions about the construction of the measurement methodology for this total effort. Requiring a standard paradigm, as opposed to having each investigation "re-invent the wheel," could also result in significant cost savings. In sum, this argues for a research program that builds on earlier work to arrive at the most scientific and cost-effective measurement methodology for job performance in the military.

V. RESEARCH PRIORITIES

As has been noted several times in this report, the purpose of this program of research is to develop a measurement methodology for job performance in the military. This research focus has guided our thinking and will shape the research priorities that are established. It has guided our development of a classification scheme and our selection of accuracy as the principal dependent measure. It is believed this measure is necessary if one is to choose with confidence the best criterion, or criteria, to validate Air Force selection and classification tests.

Having underscored the purpose and focus, what follows is a proposed systematic approach to researching the key issues in the measurement of job performance. While this section is intended to highlight needed research, it is not intended to detail specific research projects for each one listed in Section IV (Research Implications). A more detailed presentation of research issues and possible solutions is presented in Appendix A.

5.1 Measurement Methodology

Decisions about the development and use of various measures of performance, and the possible relationships between these measures, are of primary concern. Consequently, a number of measurement techniques will be evaluated in terms of their ability to measure job performance accurately. Because it is anticipated that no single available measure of job performance will accurately assess the entire criterion space, a top research priority is to identify which methods accurately measure which parts of the criterion space.

Initial efforts in this area require a priori specification of the nomological network. In addition, part of this effort should include a more detailed look at the dimensions of performance used within methods across studies, across methods within studies, and across dimensions and studies. The product of this research will be a multimethod-multidimension matrix tied to measurement of the criterion space. This line of research supports other segments of the project discussed in Section 5.3 (Research Paradigm Issues).

A related research issue of equally high importance involves empirically testing this hypothesized nomological net. Once the desired measurement methods are developed and refined, the postulated relationships can be experimentally tested. This research should take place in the later stages of the research project.

In addition to the multimethod-multidimension criterion space research, a major measurement methodology research effort will involve refinement of the WTPT technique. Because this method is still in the early stages of development, and is viewed as the benchmark and high fidelity component of the measurement process, initiation of this research is both important and urgent. Associated priority research involves the development of the performance standards scoring key to be used by personnel conducting the walk-through testing. This key is an integral part of the WTPT methodology and must be developed in conjunction with the technique itself.

With WTPT, the actual combination of tasks to be rated may be unique for each job. Existing measures of task difficulty and aptitude requirements may need to be used to equate the task ratings so individuals' scores can be compared on the same scale. Incumbents' experience ratings could be used in the same fashion.

Research should also focus on the development of alternate job performance measures. As previously noted, the walk-through testing results will serve as the reference point against which more global and less expensive measures will be compared to select the measure(s) to be used operationally.

5.2 Scale Development and Characteristics

Research on scale development and scale characteristics receives a high priority rating, not because of the amount of unsolved questions, but because of the urgency associated with scale development. Although there are well-established guidelines (as noted in Sections III and IV), a wide range of alternative job performance measures must be developed in order to compare different performance measures with the walk-through testing procedure. These should include

peer, supervisory, and self performance ratings, and should range from ratings of general performance to highly specialized, task-specific measures.

Experience ratings need to be obtained from incumbents to help moderate confounding effects. The experience measures will be particularly important to moderate the ratings, since it will not be possible to tailor the rating forms to each job incumbent. At best, the rating forms can be written for job types within specialties. In any event, the specialties should be the same AFSs as those used in the walk-through testing approach.

Research conducted during this developmental phase should focus on issues such as the number of dimensions/items required to optimize accuracy, and the degree of content overlap across dimensions, jobs, and specialties. Although this work is not a high priority on the importance continuum, the need to develop rating scales and the WTPT method in conjunction in the same AFSs elevates this research to a top priority on the urgency continuum.

5.3 Research Paradigm Issues

A major research focus must be how to operationalize and apply the accuracy criterion as various research issues are confronted. Though other criteria (e.g., reliability, psychometric effects) will be collected, the crucial issue revolves around the selection of an accuracy approach. As noted in Section IV, there are at least four main paradigms available (multitrait-multimethod analysis, videotapes, "paper people," and expected score distributions on an *a priori* basis).

Whenever possible, a combination of approaches should be used to measure accuracy. Both the *a priori* specification of expected score distributions and the *post hoc* multitrait-multimethod analyses should be used as frequently as possible because of their direct link to the hypothesized nomological net, and the eventual empirical testing of the criterion space conceptualization. However, it is believed that the best single approach to assessing the accuracy of the measuring devices is the videotape approach.

This method affords several advantages. First, because this approach is based on the development of scripts depicting varying levels of performance on different dimensions, the level of performance can be easily manipulated (as can variables such as environmental setting, sex/age of ratee, type of task being viewed, etc.). Also, normative true scores will be generated, and thus, it will be known on an *a priori* basis exactly where on the scale a rater should be rating. In addition, the level of specificity (i.e., task, job, AFS) of each tape can be varied depending on the purpose and focus of measurement.

Videotapes should be developed for a number of AFSs, using Air Force personnel or professional actors portraying Air Force personnel. Every effort should be made to make the tapes as similar to on-the-job conditions as possible. Once developed and validated, data collection will be accomplished in the field using actual military raters.

The use of videotaped vignettes will also be beneficial in answering other important performance measurement research issues. For example, in deciding on the type of training to give observers/raters of behavior, the videotapes can provide the standardized performance against which to judge the effectiveness of training. Also, videotapes of individuals being evaluated in a WTPT situation would provide an excellent mechanism for giving observation training to test administrators. Other specific issues that might be addressed include: (a) the amount of observable behavior required before a rater/observer can make an *accurate* decision about level of performance, (b) how a rater's prior knowledge of the ratee (job-related or not) affects the accuracy of ratings, and (c) the best number of dimensions or items to be used with a

particular measurement method in order to optimize the ability of the rater/observer to rate accurately.

Finally, the a priori specification of expected score distributions and the multitrait-multimethod construct validity analyses should also be included whenever possible, to gain additional insight into the accuracy of measurement. Traditional psychometric measures should also be included in any data collection effort. However, the videotape approach to measurement accuracy is considered a cornerstone of this entire project and as such, is rated high on both urgency and importance.

5.4 Identification of Possible Sources of Error Variance

This section represents an attempt to "pull together" research questions whose specific focus is the identification of possible sources of error variance. Most of these issues are rated no more than average on either the urgency or importance continuum.

A large part of the discussion in Section IV that dealt with variables such as rater characteristics, rater/ratee relationships, social context, non-work variables, performance constraints, and intervening variables, focused on issues related to error variance. Therefore, only a few issues of importance will be discussed here.

For example, one issue of concern involves the purpose of our measurement -- validation research. Because the purpose for which ratings are to be collected may affect the degree to which a rater is willing to provide accurate ratings, it may be that raters will perceive validation research as having no negative consequences for them, and therefore, provide accurate ratings. Thus, the variable of concern becomes one of the rater's trust in the uses, consequences, and benefits of the data collection effort.

Another issue of concern involves the amount of behavior observed and how this affects a rater's ability to rate accurately. How much behavior must be observed before one can be confident that the ratings are relatively accurate? Is there a point of diminishing returns? These types of questions must be answered so that the rating scales developed, raters chosen, and the amount of time spent gathering information all contribute to the accuracy of the measurement.

Many other potential sources of error variance must also be accounted for, including rater/ratee age, sex, and race congruence; the effects of various performance constraints on both raters and ratees; and the effects of non-work variables on job performance. These represent only a subset of the questions to be answered.

5.5 The Control of Error Variance

After identifying sources of error variance, research must deal with the control or elimination of at least some sources of error variance. Although a number of contributors to error variance will be controlled through standardizing procedures (randomization, equating, etc.), a major research effort should be undertaken under the general heading of rater/observer training. For instance, a training program (with a "public relations" focus) could be developed that is aimed at increasing the accuracy of ratings by increasing the raters' motivation and trust in the appraisal process.

A much larger training effort should be initiated by the second or third year of this project, directed toward training raters/observers in ways that will increase the accuracy of their evaluations. Training to improve observational skills would seem most beneficial for WTPT,

while other types of training may be required for the supervisory, peer, and self rating methods. Small laboratory and/or field pilot studies will be required to make decisions regarding length, type, and content of training prior to implementation.

Training is one of the major techniques that will be introduced to reduce error variance. In addition, instructions to observers/raters, packaging of the forms, and public relations efforts should be used. In terms of the scope of this project, training is important, but relatively less urgent than identifying sources of error variance and developing the necessary measurement methods.

5.6 Final Comments

Much of the initial research suggested here should also be repeated/refined as additional Air Force specialties are incorporated into the research effort, particularly in the areas of scale development, WTPT development, and development of the videotapes. As information and knowledge are gained from initial work in these areas, it is anticipated that time required for development will be significantly reduced.

In this program, the WTPT technique has been designated as the benchmark, high fidelity method. Consequently, much time and effort will need to be focused on this technique in order to close the credibility gap between actual and ultimate criteria.

The choice of accuracy as our measurement quality criterion is another major innovative approach that characterizes this program of research. This approach is not typical of past research efforts in performance measurement, yet recent research findings have begun to raise questions concerning the adequacy of the more traditional criteria of measurement quality.

Finally, the authors consider it essential that a job performance measurement system development effort of this magnitude be undertaken in a systematic manner. The worth of the system developed may be ultimately determined by the willingness of project personnel to systematically plan, review, and evaluate research priorities and directions during the course of this program of research.

REFERENCES

Adams, J. S. (1965). Inequity in social exchange. In L. Berkowitz (Ed.), Advances in experimental social psychology. New York: Academic Press.

Alewine, T. (1982). Performance appraisal and performance standards. Personnel Journal, 61, 210-213.

Arvey, R. D., & Hoyle, J. C. (1974). A Guttman approach to the development of behaviorally based rating scales for systems analysts and programmer/analysts. Journal of Applied Psychology, 59, 61-68.

Baird, L. (1977). Self and superior ratings of performance: As related to self esteem and satisfaction with supervision. Academy of Management Journal, 20, 291-300.

Barrett, R. (1964). The influence of the supervisor's requirements on ratings. Personnel Psychology, 17, 375-387.

Bass, A. R., & Turner, J. N. (1973). Ethnic group differences in relationships among criteria of job performance. Journal of Applied Psychology, 57, 101-109.

Bass, B. M. (1981). Stogdill's handbook of leadership: A survey of theory and research (rev. and expanded ed.). New York: Free Press.

Basset, G., & Meyer, H. (1968). Performance appraisal based on self-review. Personnel Psychology, 21, 421-430.

Bazerman, M. H., Beekun, R. I., & Schoorman, F. D. (1982). Performance evaluation in a dynamic context: A laboratory study of the impact of a prior commitment to the ratee. Journal of Applied Psychology, 67, 873-976.

Beatty, R. W., Schneier, S. E., & Beatty, J. R. (1977). An empirical investigation of perceptions of ratee behavior, frequency, and ratee behavior change using behavioral, frequency, and ratee behavior change using behavior expectation scales. Personnel Psychology, 33, 647-658.

Beer, M., Ruh, R., Dawson, J. A., McCaa, B. B., & Kavanagh, M. J. (1978). A performance management system: Research, design, introduction, and evaluation. Personnel Psychology, 31, 505-535.

Bernardin, H. J. (1977). BES vs. summated scales: A fairer comparison. Journal of Applied Pyschology, 62, 422-427.

Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. Journal of Applied Psychology, 63, 301-308.

Bernardin, H. J., Alvares, K. M., & Cranny, C. J. (1976). A recomparison of behavioral expectation scales to summated scales. Journal of Applied Psychology, 61, 564-570.

Bernardin, H. J., & Cardy, R. L. (1981). Cognitive complexity in performance appraisal: It makes no nevermind. Proceedings of the 41st Annual Meeting of the Academy of Management, 16, 306-310.

Bernardin, H. J., & Cardy, R. L. (1982). Appraisal accuracy: The ability and motivation to remember the past. Public Personnel Management Journal, 119, 352-357.

Bernardin, H. J., Cardy, R. L., & Carlyle, J. J. (1982). Cognitive complexity and appraisal effectiveness: Back to the drawing board? Journal of Applied Psychology, 67, 151-160.

Bernardin, H. J., & Kane, J. S. (1980). A second look at Behavioral Observation Scales. Personnel Pscychology, 33, 809-814.

Bernardin, H. J., Orban, J. A., & Carlyle, J. J. (1981). Performance ratings as a function of trust in appraisal, purpose for appraisal, and rater individual differences. Proceedings of the 41st Annual Meeting of the Academy of Management.

Bernardin, H. J., & Pence, E. C. (1980). Rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65, 60-66.

Bernardin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored rating scales. Journal of Applied Psychology, 66, 458-463.

Bernardin, H. J., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 62, 64-69.

Bigoness, W. J. (1976). Effect of applicant's sex, race, and performance on employer's performance ratings: Some additional findings. Journal of Applied Psychology, 61, 80-84.

Blackburn, R., & Clark, M. (1975). An assessment of faculty performance: Some correlates between administrator, colleague, student, and self-ratings. Sociology of Education, 48, 242-256.

Boehm, V. R. (1982). Assessment centers and management development. In K. M. Rowland & G. R. Ferris (Eds.), Personnel Management. Boston: Allyn & Bacon.

Borman, W. C. (1974). The rating of individuals in organizations: An alternative approach. Organizational Behavior and Human Performance, 12, 205-214.

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 60, 556-560.

Borman, W. C. (1977). Consistency of rating accuracy and rater errors in the judgment of human performance. Organizational Behavior and Human Performance, 20, 238-252.

Borman, W. C. (1978). Exploring the upper limits of reliability and validity in job performance ratings. Journal of Applied Psychology, 63, 135-144.

Borman, W. C. (1979a). Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology, 64, 410-421.

Borman, W. C. (1979b). Individual differences correlates of accuracy in evaluating others' performance effectiveness. Applied Psychological Measurement, 3, 103-115.

Borman, W. C. (1980). Performance judgments: The quest for accuracy in ratings of performance effectiveness. Paper presented at the First Annual Scientist-Practitioner Conference in Industrial-Organizational Psychology, Old Dominion University.

Borman, W. C. (1981). Evaluating the validity of the Bennett Mechanical Comprehension Test for predicting performance in operating and maintenance jobs at Florida Power and Light. Minneapolis: Personnel Decisions Research Institute.

Borman, W. C., & Dunnette, M. D. (1975). Behavior based versus trait oriented performance ratings: An empirical study. Journal of Applied Psychology, 60, 561-565.

Borman, W. C., Hough, L., & Dunnette, M. (1976). Performance ratings: An investigation of reliability, accuracy, and relationships between individual differences and rater error. Minneapolis: Personnel Decisions, Inc.

Borman, W. C., Mendel, R. M., Lammlein, S. E., & Rosse, R. L. (1981). Developing and evaluating the validity of a test battery to predict performance in transmission and distribution jobs at Florida Power and Light. Minneapolis: Personnel Decisions Research Institute.

Brown, E. M. (1968). Influence of training, method, and relationship on the halo effect. Journal of Applied Psychology, 52, 195-199.

Burnaska, R., & Hollman, T. D. (1974). An empirical comparison of the relative effects of rater response biases on three rating scale formats. Journal of Applied Psychology, 59, 307-312.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. (1973). The development and evaluation of behaviorally based rating scales. Journal of Applied Psychology, 57, 15-22.

Cascio, W. F., & Bernardin, H. J. (1981). Implications of performance appraisal litigation for personnel decisions. Personnel Psychology, 34, 211-226.

Cascio, W. F., & Phillips, N. F. (1979). Performance testing: A rose among the thorns! Personnel Psychology, 32, 751-766.

Cascio, W. F., & Valenzi, E. (1977). BARS: Effects of education and job experience of raters and ratees. Journal of Applied Psychology, 62, 278-282.

Cascio, W. F., & Valenzi, E. R. (1978). Relations among criteria of police performance. Journal of Applied Psychology, 63, 22-28.

Chapanis, A. (1976). Engineering psychology. In M. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally.

Christal, R. E. (1974, January). The United States Air Force occupational research project (AFHRL-TR-73-75, AD-774 574). Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory.

Cleveland, J., & Landy, F. (1981). The influence of rater and ratee age on two performance judgments. Personnel Psychology, 34, 19-29.

Cornelius, E., Hakel, M., & Sackett, P. (1979). A methodological approach to job classification for performance appraisal purposes. Personnel Psychology, 32, 283-297.

Cronbach, L. J., Gleser, G., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Cummings, L. L., & Schwab, D. (1973). Performance in organizations: Determinants and appraisals. Glenview, IL: Scott, Foresman.

DeCotiis, T. A. (1977). An analysis of the external validity and applied relevance of three rating formats. Organizational Behavior and Human Performance, 19, 247-266.

DeCotiis, T., & Petit, A. (1978). The performance appraisal process: A model and some testable propositions. Academy of Management Review, 3, 635-646.

Dickinson, T., & Tice, T. (1977). The discriminant validity of scales developed by retranslation. Personnel Psychology, 30, 217-227.

Dipboye, R., & dePontbriand, R. (1981). Correlates of employee reactions to performance appraisals and appraisal systems. Journal of Applied Psychology, 66, 248-251.

Downey, R., Medland, F. F., & Yates, L. G. (1976). Evaluation of peer rating system for predicting subsequent promotion of senior military officers. Journal of Applied Psychology, 61, 206-209.

Drory, A., & Ben-Porat, A. (1980). Leadership style and leniency bias in the evaluation of employees' performance. Psychological Reports, 46, 735-739.

Farr, J. L., O'Leary, B. S., & Bartlett, C. J. (1971). Ethnic group membership as a moderator of the prediction of job performance. Personnel Psychology, 24, 609-636.

Fay, C. H., & Latham, G. P. (1982). Effects of training and rating scales on rating errors. Personnel Psychology, 35, 105-116.

Feild, H., & Holley, W. (1975). Traits in performance ratings: Their importance in public employment. Public Personnel Management, 4, 327-330.

Feild, H., & Holley, W. (1977). Subordinates' characteristics, supervisors' ratings, and decisions to discuss appraisal results. Academy of Management Journal, 20, 315-321.

Feldman, J. (1981). Beyond attribution theory: Cognitive processes in performance apparaisal. Journal of Applied Psychology, 66, 127-148.

Feldman, J., & Hilterman, R. (1977). Sources of bias in performance evaluation: Two experiments. International Journal of Intercultural Relations, 1, 35-57.

Finley, D. M., Osburn, H. G., Dubin, J. A., & Jeanneret, P. R. (1977). Behaviorally based rating scales: Effects of specific anchors and disguised scale continua. Personnel Psychology, 30, 659-669.

Finn, R. (1972). Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. Vocational and Psychological Measurement, 32, 255-265.

Fiske, D. W., & Cox, J. A. (1960). The consistency of ratings by peers. Journal of Applied Psychology, 44, 11-17.

Freeberg, N. (1968). Relevance of rater-ratee acquaintance in the validity and reliability of ratings. Journal of Applied Psychology, 53, 518-524.

Friedman, B., & Cornelius, E. (1976). Effect of rater participation in scale construction of the psychometric characteristics of two rating scale formats. Journal of Applied Psychology, 61, 210-216.

Githens, W., & Elster, R. (1973). Development of a man-to-man rating scale for evaluating performance. Proceedings of the 81st Annual Convention of the American Psychological Association (pp. 757-758).

Gordon, M. (1972). An examination of the relationship between the accuracy and favorability of ratings. Journal of Applied Psychology, 56, 49-53.

Graen, G. (1976). Role-making processes within complex organizations. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally.

Greenhaus, J. H., & Gavin, J. F. (1972). The relationship between expectancies and job behavior for white and black employees. Personnel Psychology, 25, 449-455.

Grey, R. J., & Kipnis, D. (1976). Untangling the performance appraisal dilemma: The influence of perceived organizational context on evaluative processes. Journal of Applied Psychology, 61, 329-335.

Griffiths, R. (1975). The accuracy and correlates of psychiatric patients' self-assessment of their work behavior. British Journal of Social and Clinical Psychology, 14, 181-189.

Hackman, J. R. (1976). Group influences on individuals. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally.

Hakel, M. D. (1974). Normative personality factors recovered from ratings of personality descriptors: The beholder's eye. Personnel Psychology, 27, 409-421.

Hakel, M. D. (1980, April). An appraisal of performance appraisal: Sniping with a shotgun. Discussant's comments presented at the First Annual Scientist-Practitioner Conference in Industrial-Organization Psychology, Old Dominion University.

Hakel, M. D. (1982). Employment interviewing. In K. M. Rowland & G. R. Ferris (Eds.), Personnel management. Boston: Allyn & Bacon.

Hall, F., & Hall, D. (1976). Effects of job incumbents' race and sex on evaluations of managerial performance. Academy of Management Journal, 19, 476-481.

Hamel, K. (1981). Causal modeling of job satisfaction in dual career couples. Paper presented at the Fourth Annual Symposium of Applied Behavioral Science, Blacksburg, VA.

Hamner, W., Kim, J., Baird, L., & Bigoness, W. (1974). Race and sex as determinants of ratings by potential employers in a simulated work-sampling task. Journal of Applied Psychology, 59, 705-711.

Hedge, J. W. (1982). Improving the accuracy of performance evaluations: A comparison of three methods of performance appraisal training. Unpublished doctoral dissertation, Old Dominion University.

Hedge, J. W., & Kavanagh, M. J. (In press). Improving the accuracy of performance evaluations: A comparison of three methods of performance appraisal training. Journal of Applied Psychology.

Heneman, H. (1974). Comparison of self and superior ratings of managerial performance. Journal of Applied Psychology, 59, 638-643.

Hobson, C., Mendel, R., & Gibson, F. (1981). Clarifying performance appraisal criteria. Organizational Behavior and Human Performance, 28, 164-188.

Holzback, R. L. (1978). Rater bias in performance ratings: Superior, self and peer ratings. Journal of Applied Psychology, 63, 579-588.

Huck, J. R., & Bray, D. W. (1976). Management assessment center evaluations and subsequent job performance of white and black females. Personnel Psychology, 29, 13-30.

Ilgen, D., Peterson, R., Martin, B., & Boeschen, D. (1981). Supervisor and subordinate reactions to performance appraisal sessions. Organizational Behavior and Human Performance, 28, 311-330.

Ivancevich, J. M. (1979). A longitudinal study of the effects of rater training on psychometric errors in ratings. Journal of Applied Psychology, 64, 502-508.

Ivancevich, J. M. (1980). A longitudinal study of behavior expectation scales: Attitudes and performance. Journal of Applied Psychology, 65, 135-146.

Jackson, S. E., & Zedeck, S. (1982). Explaining performance variability: Contributions of goal setting, task characteristics, and evaluative contexts. Journal of Applied Psychology, 67, 759-768.

Jacobs, R., Kafry, D., & Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. Personnel Psychology, 33, 595-640.

Kane, J. S., & Bernardin, H. J. (1982). Behavioral observation scales and the evaluation of performance appraisal effectiveness. Personnel Psychology, 35, 635-641.

Kane, J. S., & Lawler, E. E., III. (1979). Performance appraisal effectiveness: Its assessment and determinants. In B. Staw (Ed.), Research in organizational behavior (Vol. 1). Greenwich, CT: JAI Press.

Kaufman, G. G., & Johnson, J. C. (1974). Scaling peer ratings: An examination of the differential validities of positive and negative nominations. Journal of Applied Psychology, 59, 302-306.

Kavanagh, M. J. (1971). The content issue in performance appraisal: A review. Personnel Psychology, 24, 653-688.

Kavanagh, M. J. (1979, May). Performance assessment in organizations: Some non-random observations. Paper presented at the AFHRL Symposium on Performance Appraisal, Brooks AFB, TX.

Kavanagh, M. J. (1980, April). Criteria for the evaluation of performance measurement techniques and performance systems. Paper presented at the First Annual Scientist-Practitioner Conference in Industrial-Organizational Psychology, Old Dominion University.

Kavanagh, M. J. (1982a). A conceptual model of job-related stress and its relation to accidents. Paper presented at the Third Annual Scientist-Practitioner Conference in Industrial-Organizational Psychology, Old Dominion University.

Kavanagh, M. J. (1982b). Evaluating performance. In K. M. Rowland & G. R. Ferris (Eds.), Personnel Management. Boston: Allyn & Bacon.

Kavanagh, M. J., DeBiasi, G., Hedge, J. W., & Miller, S. (1983, August). An empirically-based, multi-criteria evaluation of a performance evaluation system. Symposium presented at the annual meetings of the Academy of Management, Dallas.

Kavanagh, M. J., & Duffy, J. F. (1978). An extension and field test of the retranslation method for developing rating scales. Personnel Psychology, 31, 461-470.

Kavanagh, M. J., & Hedge, J. W. (1983, May). A closer look at the correlates of performance appraisal system acceptability. Paper presented at the annual meeting of the Eastern Academy of Management, Pittsburg.

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analyses of ratings. Psychological Bulletin, 75, 34-49.

Kavanagh, M. J., Vance, R., & Wright, S. (1982). The relationships between psychometric effects and accuracy for both field and laboratory data. Unpublished manuscript. State University of New York at Albany.

Keaveny, T. J., & McGann, A. F. (1975). A comparison of behavioral expectation scales and graphic rating scales. Journal of Applied Psychology, 60, 695-703.

King, L., Hunter, J., & Schmidt, F. (1980). Halo in a multidimensional force-choice performance evaluation scale. Journal of Applied Psychology, 65, 507-516.

Kingstrom, R., & Bass, A. (1981). A critical analysis of studies comparing BARS and other rating formats. Personnel Psychology, 34, 263-289.

Kirby, P. (1981). Part 1: A systematic approach to performance appraisal. Management World, 10, 16-17, 28.

Klimoski, R. J., & London, M. (1974). Role of the rater in performance appraisal. Journal of Applied Psychology, 59, 445-451.

Knowlton, W. A., & Mitchell, T. R. (1980). Effects of causal attribution on a supervisor's evaluation of subordinate performance. Journal of Applied Psychology, 65, 459-466.

Kraut, A. I. (1975). Prediction of managerial success by peer and training staff ratings. Journal of Applied Psychology, 60, 14-19.

Lahey, M., & Saal, F. (1981). Evidence incompatible with a cognitive compatibility theory of rating behavior. Journal of Applied Psychology, 66, 706-715.

Landy, F., Barnes-Farrell, J., & Cleveland, J. (1980). Perceived fairness and accuracy of performance evaluation: A follow-up. Journal of Applied Psychology, 65, 355-356.

Landy, F., Barnes, J., & Murphy, K. (1978). Correlates of perceived fairness and accuracy of performance evaluation. Journal of Applied Psychology, 63, 751-754.

Landy, F. J., & Farr, J. L. (1980). Performance ratings. Psychological Bulletin, 87, 72-107.

Latham, G. P., Saari, L. M., & Fay, C. (1980). BOS, BES, and baloney: Raising Kane with Bernardin. Personnel Psychology, 33, 815-821.

Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales of performance appraisal purposes. Personnel Psychology, 30, 225-268.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 60, 550-555.

Lawler, E. E., III. (1967). The multitrait-multirater approach to measuring managerial job performance. Journal of Applied Psychology, 51, 369-381.

Lawler, E. E., III. (1976). Control systems in organizatons. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally.

Lee, D., & Alvares, K. (1977). Effects of sex on descriptions and evaluations of supervisory behavior in a simulated industrial setting. Journal of Applied Psychology, 62, 405-410.

Levine, J., & Butler, J. (1952). Lecture vs. group discussion in changing behavior. Journal of Applied Psychology, 36, 29-33.

Lewin, A. Y., & Zwany, A. (1976). Peer nominations: A model, literature critique, and a paradigm for research. Personnel Psychology, 29, 423-447.

London, M., & Poplawski, J.R. (1976). Effects of information on stereotype development in performance appraisal and interview contexts. Journal of Applied Psychology, 61, 199-205.

Maas, J. (1965). Patterned scale expectation interview: Reliability studies on a new technique. Journal of Applied Psychology, 49, 431-433.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. Journal of Applied Psychology, 67, 280-296.

MacKinney, A. C. (1967). The assessment of performance change: An inductive example. Organizational Behavior and Human Performance, 2, 56-72.

Massey, R., Mullins, C., & Earles, J. (1978). Performance appraisal ratings: The content issue (AFHRL-TR-78-69, AD-A064 690). Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

McAfee, B., & Green, B. (1977). Selecting a performance appraisal method. The Personnel Administrator, 76, 61-64.

McCall, M. W., & DeVries, D. L. (1976). Appraisal in context: Clashing with organizatonal realities. Paper presented at the 84th Annual Convention of the American Psychological Association, Washington, DC.

Meyer, H. (1980). Self-appraisals of job performance. Personnel Psychology, 33, 291-295.

Mitchell, T. R., & Liden, R. C. (1982). The effects of social context on performance evaluations. Organizational Behavior and Human Performance, 29, 241-256.

Mobley, W. (1982). Supervisor and employee race and sex effects on performance appraisal: A field study of adverse impact and generalizability. Academy of Management Journal, 25, 598-606.

Morano, R. (1979). An Rx for performance appraisal. Personnel Journal, 58, 306-307.

Moses, J. L., & Boehm, V. (1975). Relationship of assessment center performance to management progress of women. Journal of Applied Psychology, 60, 527-529.

Muczyk, J., & Gable, M. (1981). Unidimensional (global) vs. multidimensional composite performance appraisals of store managers. Journal of the Academy of Marketing Science, 9, 191-205.

Mullins, C. J., & Force, R. C. (1962). Rater accuracy as a generalized ability. Journal of Applied Psychology, 46, 191-193.

Mullins, C. J., Seidling, K., Wilbourn, J. M., & Earles, J. (1979). Rater accuracy study. (AFHRL-TR-78-89, AD-A066 779). Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Mullins, C. J., Weeks, J. L., & Wilbourn, J. M. (1978). Ipsative rankings as an indicator of job-worker match (AFHRL-TR-78-70, AD-A065 053). Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Murphy, K. R. (1982). Difficulties in the statistical control of halo. Journal of Applied Psychology, 67, 161-164.

Murphy, K. R., Garcia, M., Kerkar, S., Martin, C., & Balzar, W. K. (1982). Relationship between observational accuracy and accuracy in evaluating performance. Journal of Applied Psychology, 67, 320-325.

Murphy, K. R., Martin, C., & Garcia, M. (1982). Do behavioral observation scales measure observation? Journal of Applied Psychology, 67, 562-567.

Naylor, J. C., Pritchard, R. D., & Ilgen, D. R. (1980). A theory of behavior in organizations. New York: Academy Press.

Nieva, V., & Gutek, B. Sex effects on evaluation. (1980). Academy of Management Review, 5, 167-276.

Nugent, W. A., Laabs, G. J., & Panell, R. C. (1982). Performance test objectivity: A comparison of rater accuracy and reliability using three observation forms (NPRDC-TR-82-30). San Diego, CA: Navy Personnel Research and Development Center.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw Hill.

Owens, W. A., & Champagne, J. A. (1965). A selected bibliography on biographical data. Greensboro, NC: The Creativity Research Institute of the Richardson Foundation.

Payne, R., & Pugh, D. S. (1976). Organizational structure and climate. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago, Rand McNally.

Peters, L., Fisher, C., & O'Connor, E. (1982). The moderating effect of situational control of performance variance on the relationship between individual differences and performance. Personnel Psychology, 35, 609-621.

Peters, L., & O'Connor, E. (1980). Situational constraints and work outcomes: The influence of a frequently overlooked construct. Academy of Management Review, 5, 391-397.

Peters, L., & O'Connor, E., & Rudolf, C. (1980). The behavioral and affective consequences of performance-relevant situational variables. Organizational Behavior and Human Performance, 25, 79-96.

Quinn, J. (1969). Bias in performance appraisal. Personnel Administrator, 16, 40-43.

Rizzo, W., & Frank, F. (1977). Influence of irrelevant cues and alternate forms of graphic rating scales on the halo effect. Personnel Psychology, 30, 405-417.

Robertson, I., & Downs, S. (1979). Learning and the prediction of performance: Development of trainability testing in the United Kingdom. Journal of Applied Psychology, 64, 42-50.

Ronan, W. W., & Prien, E. P. (1971). Perspective on the measurement of human performance. New York: Appleton-Century-Crofts.

Rose, R. M., Jenkins, C. D., & Hurst, M. W. (1978). Air traffic controller health change study: A prospective investigation of physical, psychological, and work-related changes. Report to the Federal Aviation Administration, Boston University School of Medicine.

Rosinger, G., Myers, L., Leve, G., Loar, M., Mohrman, S., & Stock, J. (1982). Development of a behaviorally based performance appraisal system. Personnel Psychology, 35, 75-88.

Saal, F. E., & Landy, F. J. (1977). The mixed standard rating scale: An evaluation. Organizational Behavior and Human Performance, 18, 19-35.

Sauser, W. I., & Pond, S. B. (1981). Effects of rater training and participation on cognitive complexity: An exploration of Schneier's cognitive reinterpretation. Personnel Psychology, 34, 563-577.

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. Personnel Psychology, 24, 419-434.

Schneier, C. (1977). Multiple rater groups and performance appraisal. Public Personnel Management, 6, 13-20.

Schneier, C., & Beatty, R. (1978). The influence of role prescriptions on the performance appraisal process. Academy of Management Journal, 21, 129-135.

Schwab, D. P., Heneman, H. G., III. (1978). Age stereotyping in performance appraisal. Journal of Applied Psychology, 63, 573-578.

Schwab, D. P., & Heneman, H. G., & DeCotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 28, 549-562.

Schwartz, D. (1977). A job sampling approach to merit system examining. Personnel Psychology, 30, 175-185.

Scott, W. E., & Hamner, W. C. (1975). The influence of variation in performance profiles on the performance evaluation process: An examination of the validity of the criterion. Organizational Behavior and Human Performance, 14, 360-370.

Seaton, R. (1974). Why ratings are better than comparisons. Journal of Advertising Research, 14, 45-48.

Shapira, Z., & Shirom, A. (1980). New issues in the use of behaviorally anchored rating scales: Level of analysis, the effects of incidence frequency, and external validation. Journal of Applied Psychology, 65, 517-523.

Sharit, J., & Salvendy, G. (1982). Occupational stress: Review and reappraisal. Human Factors, 24, 129-162.

Siegel, L. (1982). Paired comparison evaluations of managerial effectiveness by peers and supervisors. Personnel Psychology, 35, 843-852.

Smith, P. C. (1976). Behaviors, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago, Rand McNally.

Spool, M. D. (1978). Training programs for observers of behavior: A review. Personnel Psychology, 31, 853-888.

Steers, R. M., & Porter, L. W. (1979). Motivation and Work Behavior (2nd ed.). New York: McGraw Hill.

Stockford, L., & Bissell, H. W. (1949). Factors involved in establishing a management rating scale. Personnel, 26, 94-116.

Stone, T. (1970, October). Sources of evaluator bias in performance appraisal. Experimental Publication System, 8, Ms. #290-12, 1-10.

Taft, R. (1971). The ability to judge people. In W. W. Ronan & E. P. Prien (Eds.), Perspectives on the measurement of human performance. New York: Appleton-Century-Crofts.

Taylor, E. K., & Hastman, R. (1956). Relation of format and administration to the characteristics of graphic rating scales. Personnel Psychology, 9, 181-206.

Taylor, R. L., & Wilsted, W. D. (1974). Capturing judgment policies: Field study of performance appraisal. Academy of Management Journal, 17, 440-449.

Thornton, G. C. (1980). Psychometric properties of self-appraisals of job performance. Personnel Psychology, 33, 363-371.

Thornton, G. C., & Zorich, S. (1980). Training to improve observer accuracy. Journal of Applied Psychology, 65, 351-354.

Tosti, D. (1979). Performance measures for job certification and system validation. Training and Development Journal, 33, 20-23.

Turnage, J., & Muchinsky, P. M. (1982). Transsituational variability in human performance within assessment centers. Organizational Behavior and Human Performance, 30, 174-200.

Vineberg, R., & Joyner, J. (1982, March). Prediction of job performance: Review of military studies (NPRDC-TR-82-37). San Diego: Naval Personnel Research and Development Center.

Vroom, V. H. (1976). Leadership. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally.

Warmke D., & Billings, R. S. (1979). A comparison of training methods for improving the psychometric quality of experimental and administrative performance ratings. Journal of Applied Psychology, 64, 124-131.

Wendelken, D., & Inn, A. (1981). Nonperformance influences on performance evaluations. Journal of Applied Psychology, 66, 149-158.

Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of ratings. Personnel Psychology, 35, 521-551.

Wiley, L. N., & Hahn, C. P. (1977, December). Task level job performance criteria development (AFHRL-TR-77-75, AD-A055 694), Brooks AFB, TX: Occupation and Manpower Research Division.

Wilkinson, R. (1969). Some factors influencing the effect of environmental stressors on performance. Psychology Bulletin, 72, 260-272.

Williams, W., & Seiler, D. (1973). Supervisor and subordinate participation in the development of behaviorally anchored rating scales. Journal of Industrial and Organizational Psychology, 1, 1-12.

Wood, R., & Mitchell, T. (1981). Manager behavior in a social context: The impact of impression management on attributions and disciplinary actions. Organizational Behavior and Human Performance, 28, 356-378.

Yukl, G. A. (1981). Leadership in organizations. Englewood Cliffs, NJ: Prentice-Hall.

Zammuto, R. F., London, M., & Rowland, K. M. (1982). Organization and rater differences in performance appraisals. Personnel Psychology, 35, 643-658.

Zedeck, S., & Cascio, W. F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. Journal of Applied Psychology, 67, 752-758.

Zedeck, S., Imparato, N., Drausz, M., & Oleno, T. (1974). Development of behaviorally anchored rating scales as a function of organizational level. Journal of Applied Psychology, 59, 249-252.

Zedeck, S., Kafry, D. (1977). Capturing rater policies for processing evaluation data. Organizational Behavior and Human Performance, 18, 269-294.

Zedeck, S., Kafry, D., & Jacobs, R. (1976). Format and scoring variations in behavioral expectation evaluations. Organizational Behavior and Human Performance, 17, 171-184.

APPENDIX A: RESEARCH ISSUES

The specific research issues and solutions included here are categorized according to the headings identified in Section IV - Research Implications. Each research issue is rated on nine-point scales of importance and urgency (1-most important/urgent; 9-least important/urgent). In addition, estimated start and completion dates for each research effort are included to help "ground" the ratings within a 5-year R&D period.

RATINGS		YEARS	
IMPORTANCE	URGENCY	START	COMPLETION

INDIVIDUAL CHARACTERISTICS

I) Research focusing on the trust respondents/ratees have in the measurement system

A) Issues

1) Are instructions on the rating forms adequate to ensure respondent trust?	5	8	3	4
2) Is a training program necessary to ensure trust?	5	8	3	4
3) Who could/should administer such a program (e.g., trained scientists, on-site personnel)?	5	8	3	4
4) Would a public relations campaign be as effective as other means to improve trust?	5	8	3	4

B) Solutions

1) Some of this research can be tagged on to other efforts (and/or be in-house generated); e.g., questionnaire administered during early testing of ratings/WTPT procedures to determine best way to convey message that while effort is for research purposes only, it is important.

2) A more extensive effort would involve an epxerimental design manipulating type and degree of training, type of person administering tests, and then measuring trust

II) Research focusing on the individual differences between raters/observers

A) Issues

1) Can/should WTPT administrators be selected according to certain individual difference criteria? (In other words, are certain attributes/abilities predictive of observational accuracy?)	4	10	5	5

	RATINGS		YEARS	
	IMPORTANCE	URGENCY	START	COMPLETION

INDIVIDUAL CHARACTERISTICS -- continued

B) Solutions
1) Some information can be gained from existing files (ASVAB scores, etc.)
2) Additional information can be obtained by comparing accuracy of different types/groups of raters/observers

III) Research focusing on rater/observer's understanding of the job

A) Issues

1) How important is the rater/observer's understanding of the job to ensuring accuracy?	4	4	2	3
2) If this is important, can the person be trained to be more knowledgeable about the job?	4	4	2	3
3) If so, how does this relate to accuracy?	4	4	2	3

B) Solutions
1) Best way -- use developed videotapes with normative target scores to assess accuracy of raters with differing amounts of knowledge of the job (lab and/or field setting)
2) Train observers on job content, etc. (maybe using videotapes of WTPT with differing levels of proficiency, varying amount of information/length of training
a) Once again evaluating in terms of impact on accuracy (this feeds into Rater/Ratee II)

RATER/RATEE RELATIONSHIPS

I) Research focusing on degree of acquaintance between rater and ratee

A) Issues

1) What is the relationship between rater & ratee acquaintance and accuracy (does the degree of acquaintance help or hinder accuracy -- maybe it's an inverted U relationship)?	5	5	3	5
2) Related to this question, how can acquaintance/prior knowledge error variance be reduced?	5	5	3	5

RATINGS		YEARS	
IMPORTANCE	URGENCY	START	COMPLETION

RATER/RATEE RELATIONSHIPS -- continued

B) Solutions
1) Test experimentally by manipulating degree of acquaintance (determined by questionnaire) & measuring amount of error variance in ratings
2) Use videotapes & manipulate degree of acquaintance by amount of prior information presented
3) In relation to Issue #2, this may be a training question & can be tagged on to other research on reduction of error variance

II) Research focusing on amount of observable behavior required
A) Issues
1) What degree/amount of observable, relevant behavior is required (how much is necessary, and when no more helps in terms of accuracy)? 3 3 2 2

B) Solutions
1) Post hoc data analyses (e.g., regression analysis) to help determine, for instance, how many dimensions/items are required
2) Use videotapes & manipulate amount of information raters/observers receive and see how it affects accuracy
3) Just ask raters how long it took/how many dimensions, etc.

III) Research focusing on rater/ratee sources of error variance
A) Issues
1) Do sex, age, race effects (and/or interactions) exist with WTPT/rating forms? 5 5 3 5

B) Solutions
1) Manipulate/control these variables in all possible combinations (tag-on research)

	RATINGS		YEARS	
	IMPORTANCE	URGENCY	START	COMPLETION

MEASUREMENT METHODS

I) Research focusing on the relationships among measurement methods

A) Issues

1) A priori specification of the relationship between measurement methods and the dimensions of job performance within the criterion space (what method measures what piece of the criterion space; where is the overlap, etc.) — 1 | 2 | 1 | 1

2) Empirical test of the hypothesized nomological network — 2 | 10 | 5 | 5

3) The WTPT procedure is envisioned as the benchmark & high fidelity component of the measurement methods. Therefore, research concerning refinement of approach, etc. will be undertaken — 1 | 1 | 1 | 2

B) Solutions

1) Literature review of methods used, criterion space measured, etc., & theoretical development of a measurement method by dimension matrix

2) Look at uniqueness of dimensions (i.e., when we're validating ASVAB, do we need to validate it against some sort of weighted checklist, or a composite, develop a synthetic criterion, etc.?)
   a) possibly use a policy-capturing approach

C) Consequences

MUCH OF THIS TOTAL RESEARCH EFFORT REVOLVES AROUND THE DETERMINATION OF THE MOST ACCURATE MEASURES OF DIFFERENT ASPECTS OF PERFORMANCE, AND THUS, THIS RESEARCH IS CRITICAL TO THE OVERALL SUCCESS OF THE EFFORT

RATING SCALE DEVELOPMENT

I) Research focusing on various aspects of scale development

A) Issues

1) What dimensions should be used (see measurement method IA & B; scale characteristics IA & B)? — 7 | 1 | 1 | 3

2) Who should generate critical incidents? — 8 | 1 | 1 | 3

3) Who should provide scalar points? — 8 | 1 | 1 | 3

RATINGS		YEARS	
I	U	S	C
M	R	T	O
P	G	A	M
O	E	R	P
R	N	T	L
T	C		E
A	Y		T
N			I
C			O
E			N

RATING SCALE DEVELOPMENT -- continued

B) Solutions
1) See characteristics IB & methods IB

C) Consequences
FAILURE TO PURSUE THIS LINE OF RESEARCH WILL REDUCE AF'S ABILITY TO DETERMINE THE FIDELITY OF MEASUREMENT METHODS

SCALE CHARACTERISTICS

I) Research focusing on scale characteristics that may impact on accuracy

A) Issues

1) How many dimensions are necessary to solve the criterion deficiency problem? (This will be an initial effort -- it can be refined/validated later); how many dimensions/items will be required to optimize accuracy? 7 1 1 1

B) Solutions

1) Factor analysis or similar statistical procedure to determine number of dimensions

2) Later empirical tests

a) with videotapes -- manipulating number of dimensions on rating form and then measuring accuracy

b) collect field data using rating forms with differing numbers of dimensions & compare

PERFORMANCE STANDARDS

I) Research focusing on performance standards development for rating forms

A) Issues

1) In relation to the development of items/dimensions/scales, should performance standards also be developed? 8 9 4 5

B) Solutions

1) Relatively unimportant when used with "for research purposes only" paradigm -- but may be tag-on research at some point in project

RATINGS		YEARS	
IMPORTANCE	URGENCY	START	COMPLETION

PERFORMANCE STANDARDS -- continued

II) Research focusing on the development of performance standards for use with WTPT procedures

A) Issues

1) In the development of performance standards to use as guidelines to rate performance using WTPT method, the issue is: who will develop these standards/scoring keys and how will they be developed? 3 3 2 2

B) Solutions

1) Job experts should develop the performance standards/scoring keys to be used by the WTPT administrators

a) development must ensure accurate scoring of observations

C) Consequences

THE ACCURACY/FIDELITY OF THIS MEASUREMENT METHOD HINGES ON THE DEVELOPMENT OF STANDARDS -- AND, LIKEWISE, THE ABILITY TO TIE SELECTION, TRAINING, ETC. TO HANDS-ON PERFORMANCE DEPENDS ON THE ACCURACY/FIDELITY OF THIS MEASUREMENT METHOD

SOCIAL CONTEXT

I) Research focusing on influences of social context

A) Issues

1) What effects do various social context variables have on measurement accuracy (which of these variables contribute to error variance; e.g., contrast errors)? 8 9 4 5

2) Do interpersonal skills impact on judgments of technical competence for rating and WTPT methods (criterion contamination)? 2 4 2 3

B) Solutions

1) Tag-on research -- see Rater/Ratee IIIB

2) Manipulate interpersonal skills of ratee and measure influences on rater/observer

b) this can be done in field, or with videotapes

	RATINGS		YEARS	
	IMPORTANCE	URGENCY	STARTING	COMPLETION
NON-WORK VARIABLES				
I) Research focusing on non-work variables (family problems, health, etc.)				
A) Issues				
1) Do, and if so, how do non-work variables affect performance, and consequently, the way that overall performance is perceived?	9	9	4	4
B) Solutions				
1) See Rater/Ratee III B				
2) In terms of WTPT, interviewer may need to obtain info from ratee's supervisor concerning non-work variables or possibly administer a questionnaire				
PERFORMANCE CONSTRAINTS				
I) Research focusing on performance constraints (e.g., supplies, machines, etc.) that impact on individual performance	7	9	4	5
A) & B) Issues & Solutions see non-work variables				
ORGANIZATION/UNIT NORMS				
I) Research focusing on organization and unit norms that impact on rating quality				
A) Issues				
1) How to best approach the problem of selling the data collection "for research purposes only" (really a matter of breaking the organizational unit rating set)	5	8	3	4
B) Solutions				
1) see individual differences I A & B				

IMPORTANCE	RATINGS AGENCY	YEARS STAT	YEARS COMPLETION

RATER TRAINING

I) Research focusing on training raters/observers

A) Issues

1) It seems apparent that some type of rater training is necessary — 2 5 3 5

a) What kind of training is necessary in order to improve accuracy?

b) Who should do the training?

c) What should the content of training be?

d) What should the length of training be?

e) Should refresher training be used?

All of these issues should be evaluated in terms of measurement & performance accuracy

2) Observer training -- with a - e, same as above.

B) Solutions

1) Experimentally test out different types of training--content, length, etc. (manipulating each variable)

2) Experimentally test which type of training (psychometric error, observation, decision-making) improves accuracy of ratings and/or observations

a) this training research can be conducted both in the field with real ratings/observations & with videotaped vignettes for both WTPT & rating approaches.

C) Consequences

FAILURE TO PURSUE THIS LINE OF RESEARCH WILL SEVERELY REDUCE THE AIR FORCE'S ABILITY TO DETERMINE THE ACCURACY/CONSTRUCT VALIDITY OF VARIOUS CRITERION MEASURES.

INTERVENING VARIABLES

I) Research focusing on variables that intervene between independent variables and accuracy of ratings

A) Issues

1) Observation/decision heuristics were discussed in Individual Differences section I A & II A — 4 6 3 3

2) Acceptability of system by ratees/raters and motivation of raters/observers are important in terms of system accuracy — 6 8 3 4

RATINGS		YEARS	
IMPORTANCE	URGENCY	START	COMPLETION

INTERVENING VARIABLES -- continued

B) Solutions
   1) See A1
   2) Assess by means of questionnaire(s) the acceptability of the system

MEASUREMENT & RESEARCH PARADIGM ISSUES

I) Research focusing on research paradigms to use
   A) Issues
      1) Research deciding which measures should be used to assess accuracy/construct validity (multitrait/multimethod construct validity, paper people, videotape vignettes, a priori specification) — 1 1 1 3
         a) should we use more than one (we should be at least consistent/systematic to some extent)?
   B) Solutions
      1) Development of videotapes of ratee performance so as to generate normative target scores & assess accuracy
      2) In house a priori specifications of criterion space
   C) Consequences
      FAILURE TO ADOPT A SYSTEMATIC APPROACH TO MEASURING ACCURACY/CONSTRUCT VALIDITY WILL UNDERMINE THE WORTH OF THE ENTIRE PROJECT.

• GOVERNMENT PRINTING OFFICE: 1987 - [illegible]